Electronic Journal of Statistics
Vol. 11 (2017) 608–639
ISSN: 1935-7524
DOI: 10.1214/17-EJS1244

Cross-calibration of probabilistic forecasts

Christof Strähl and Johanna Ziegel

Institute of Mathematical Statistics and Actuarial Sciences,
University of Bern, Switzerland
e-mail: christof.straehl@stat.unibe.ch; johanna.ziegel@stat.unibe.ch

Abstract: When providing probabilistic forecasts for uncertain future events, it is common to strive for calibrated forecasts, that is, the predictive distribution should be compatible with the observed outcomes. Often, there are several competing forecasters of different skill. We extend common notions of calibration where each forecaster is analyzed individually, to stronger notions of cross-calibration where each forecaster is analyzed with respect to the other forecasters. In particular, cross-calibration distinguishes forecasters with respect to increasing information sets. We provide diagnostic tools and statistical tests to assess cross-calibration. The methods are illustrated in simulation examples and applied to probabilistic forecasts for inflation rates by the Bank of England. Computer code and supplementary material (Strähl and Ziegel, 2017a,b) are available online.

Keywords and phrases: Calibration, predictive distribution, prediction space, probability integral transform, proper scoring rule.

Received November 2016.

Contents

1. Introduction

In the past decades, probabilistic forecast, specifying a complete predictive probability distribution for an uncertain future event, have replaced point forecasts in a number of applications including weather forecasting, climate predictions and economics; see Gneiting and Katzfuss (2014) for a recent overview. Murphy and Winkler (1987); Gneiting et al. (2007) have formulated the guiding principle for a probabilistic forecast to "maximize sharpness subject to calibration". Calibration refers to the statistical compatibility of the forecasts and the observations. Sharpness, on the other hand, is a property that concerns the forecast only. Roughly speaking, a forecast is sharper the more concentrated the distribution is, with point forecasts as a limiting case. Gneiting et al. (2007) have formulated their principle in order to pick the "better" of two calibrated forecasts. While it is generally acknowledged that forecasts should be calibrated (Dawid, 1984; Diebold et al., 1998), it is not universally accepted that it is necessary to consider sharpness as a further criterion for forecast evaluation (Mitchell and Wallis, 2011).

In this manuscript, we propose concepts of cross-calibration in order to formalize the influence of competing forecasters amongst each other and with respect to the observations. Essentially, a cross-calibrated forecaster not only uses her own information optimally but also incorporates the information of the competing forecasters in an optimal way. Moreover, a cross-calibrated forecast is automatically the sharpest; see Section 3. The notions we propose are a broad generalization of the existing notions of calibration of Gneiting and Ranjan (2013). Furthermore, we extend their prediction space setting to allow for serial dependence which is the usual situation in forecasting applications. We are able to extend the result of Diebold et al. (1998) of uniformity and independence of probability integral transform (PIT) values to our general framework.

There is a large literature on evaluating predictive performance based on tests for uniformity and independence of PIT values; see Dawid (1984); Diebold et al. (1998); Berkowitz (2001) to mention just a few. An ideal forecast should be preferred by all stakeholders and it always leads to uniformly distributed, independent PIT values. However, if forecasters are allowed to use more information than the past realizations of the quantity of interest, uniformity and independence of PIT values only gives limited information about the quality of the forecasts. Examples of this fact can be found in Hamill (2001); Gneiting et al. (2007); see also Examples 2.5 and 2.10. Such additional information can consist of time series of covariates or expert opinion on parameters in a model. Holzmann and Eulert (2014) show that comparing forecasts with respect to proper scoring rules respects increasing information sets in the sense that a more informative forecast will be preferred. In this paper, we address the problem from a different angle and introduce notions of cross-calibration in order to account for different information sets. The advantage of our approach is that we define an optimality property of a forecast, similar to the classical notion of uniformity and independence of PIT values. When comparing forecasts using proper scoring rules, it is only possible to make comparative statements

rather than statements about optimality; see Section 3 for a discussion on how cross-calibrated forecasts are ranked by proper scoring rules. While the notions of cross-calibration appear rather technical at first sight, we propose generally applicable tests that show good power for small to moderate sample sizes. In particular, the p-value plots for the conditional exceedance tests presented in Section 6.1 allow to identify which quantiles of the predictive distributions are not optimal. This may be relevant for applications in risk management, where it has been proposed to use PIT values for backtesting (Campbell, 2005).

Notions of cross-calibration have previously been considered in the literature for binary or categorical outcomes. Al-Najjar and Weinstein (2008) consider a test which an uninformed forecaster cannot pass with high probability when an informed forecaster is present. The notion of cross-calibration by Feinberg and Stewart (2008) takes into account that several forecasters may influence each other, and the one with the largest information set should be preferred. Their test is a generalization of the calibration test suggested by Dawid (1985). We review the Feinberg and Stewart (2008) cross-calibration test in the supplementary material to this paper and illustrate it with a simulation example. In this paper, we generalize the cross-calibration notions of Feinberg and Stewart (2008) to forecasts of real valued outcomes including diagnostic tools and statistical tests to assess cross-calibration in applications. We have chosen to work in the framework of prediction spaces as introduced by Gneiting and Ranjan (2013), and extend it to allow for serial dependence.

The paper is organized as follows. In Section 2, we review and extend the notion of a prediction space and generalize the notions of calibration for individual forecasters to multiple forecasters. In Section 3, we review the decomposition for strictly proper scoring rules by Bröcker (2009) and show that a cross-calibrated forecaster is automatically the sharpest one. In Section 4, we treat the special case of binary outcomes and relate our work to the existing results of Feinberg and Stewart (2008). We introduce diagnostic tools for checking cross-calibration and illustrate their usefulness in a simulation study in Section 5. Statistical tests for cross-calibration are derived in Section 6. We analyze the Bank of England density forecasts for inflation rates in Section 7. Finally, the paper concludes with a discussion in Section 8. All proofs are deferred to the appendix. All simulations and data examples were done in R (R Core Team, 2015).

2. Notions of cross-calibration

We follow Gneiting and Ranjan (2013) by introducing the notion of a *prediction space*.

Definition 2.1 (one-period prediction space)**.** Let $k \geq 1$ be an integer. Consider a probability space $(\Omega, \mathcal{A}, \mathbb{Q})$ together with sub-σ-algebras $\mathcal{A}_1, \ldots, \mathcal{A}_k \subset \mathcal{A}$. A *one-period prediction space* is a collection of a real-valued random variable Y on $(\Omega, \mathcal{A}, \mathbb{Q})$, Markov kernels $F_i : \Omega \times \mathcal{B}(\mathbb{R}) \to [0,1]$ from $(\Omega, \mathcal{A}_i)$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ for $1 \leq i \leq k$ and a standard uniform random variable V on $(\Omega, \mathcal{A}, \mathbb{Q})$ independent of $\mathcal{A}_1, \ldots, \mathcal{A}_k$ and Y.

The integer k corresponds to the number of forecasters. The Markov kernel F_i represents forecaster i and yields for each outcome $\omega \in \Omega$ a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. The σ-algebra $\mathcal{A}_i$ can be seen as the information set available to forecaster i. The random variable Y is the observation. The random variable V is needed for technical reasons. It allows to define the probability integral transform (PIT) in Definition 2.6 below.

We term the prediction space proposed by Gneiting and Ranjan (2013) a *one-period* prediction space as it is only concerned with predictions for an outcome Y at one time point. While this framework is sufficient to define various notions of calibration and cross-calibration of forecasters in principle, a statistical analysis of calibration is only possible if we can assume that we have independent forecast-observation tuples $(F_{1,t}, \dots, F_{k,t}, Y_{t+1}, V_t)$ for $1 \leq t \leq N$. This assumption is unrealistic in most forecasting situations. Therefore, we propose to extend the prediction space setting, allowing for serial dependence as follows.

Definition 2.2 (prediction space for serial dependence)**.** Let $k \geq 1$ be an integer. Consider a probability space $(\Omega, \mathcal{A}, \mathbb{Q})$ together with filtrations $(\mathcal{A}_{1,t})_{t\in\mathbb{N}}$, $\dots, (\mathcal{A}_{k,t})_{t\in\mathbb{N}}$ with $\mathcal{A}_{i,t} \subset \mathcal{A}$ for all $1 \leq i \leq k$, $t \in \mathbb{N}$. A *prediction space for serial dependence* is a collection of a sequence of real-valued random variables $(Y_t)_{t\in\mathbb{N}}$ with the filtration $(\mathcal{T}_t)_{t\in\mathbb{N}}$ generated by $(Y_t)_{t\in\mathbb{N}}$, that is, $\mathcal{T}_t = \sigma(Y_s, s \leq t)$, a sequence of Markov kernels $(F_{i,t})_{t\in\mathbb{N}}$ for $1 \leq i \leq k$, where each $F_{i,t}$ is from $(\Omega, \sigma(\mathcal{A}_{i,t}, \mathcal{T}_t))$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and an iid sequence $(V_t)_{t\in\mathbb{N}}$ of standard uniform random variables that is independent of everything else.

While the formalism in Definition 2.2 is fairly complicated, we would like to emphasize that it reflects a common situation in applications. The notation in Definition 2.2 is chosen such that $\mathcal{A}_{i,t}$ encodes the information of the i-th forecaster $F_{i,t}$ at time t to predict the outcome Y_{t+1} at the next time point. Additionally, all forecasters $F_{1,t}, \dots, F_{k,t}$ have access to the past realizations of Y_t in principle, that is, to the information contained in $\mathcal{T}_t$. This means, we have separated the information of forecaster i into two parts, the information of past realizations of the outcome $\mathcal{T}_t$, that is available to all forecasters, and a personal information set $\mathcal{A}_{i,t}$ that she acquires from other sources.

All further statements are within the prediction space for serial dependence and expressions such as *almost surely* are with respect to the probability measure $\mathbb{Q}$. In the prediction space for serial dependence, $F_{i,t}$ is termed *ideal* with respect to $\mathcal{A}_{i,t}$ if

$$F_{i,t} = \mathcal{L}(Y_{t+1} | \mathcal{A}_{i,t}, \mathcal{T}_t) \quad \text{almost surely,}$$

where $\mathcal{L}(X \mid \mathcal{G})$ denotes the conditional law of a random variable X with respect to the σ-algebra generated by $\mathcal{G}$. In the case of independent forecast-observation tuples, we recover the definition of an ideal forecaster of Gneiting and Ranjan (2013), that is, in the one-period prediction space setting, F_i is *ideal* with respect to $\mathcal{A}_i$ if

$$F_i = \mathcal{L}(Y | \mathcal{A}_i) \quad \text{almost surely;}$$

see also Tsyplakov (2011, 2013). We generalize this notion as follows.

Definition 2.3 (cross-ideal)**.** In the prediction space setting for serial dependence, we call $F_{i,t}$ *cross-ideal* with respect to $\mathcal{A}_{1,t}, \dots, \mathcal{A}_{k,t}$ if

$$F_{i,t} = \mathcal{L}(Y_{t+1} | \mathcal{A}_{1,t}, \dots, \mathcal{A}_{k,t}, \mathcal{T}_t) \quad \text{almost surely.} \tag{1}$$

A cross-ideal forecaster does not only use her own information optimally but also the information available to the other forecasters. In fact, at time t, her information $\mathcal{A}_{i,t}$ contains all relevant information of all the forecasters because $F_{i,t}$ is a version of the conditional distribution of Y_{t+1} with respect to $\sigma(\mathcal{A}_{1,t}, \dots, \mathcal{A}_{k,t}, \mathcal{T}_t)$. On the other hand, $F_{i,t}(\cdot, B)$ is $\sigma(\mathcal{A}_{i,t}, \mathcal{T}_t)$-measurable for all $B \in \mathcal{B}(\mathbb{R})$ by Definition 2.2 of the prediction space for serial dependence. Thus, $F_{i,t}$ is also a version of the conditional distribution of Y_{t+1} with respect to $\sigma(\mathcal{A}_{i,t}, \mathcal{T}_t)$ and hence $F_{i,t} = \mathcal{L}(Y_{t+1} | \mathcal{A}_{i,t}, \mathcal{T}_t)$ almost surely. Therefore, each cross-ideal forecaster is ideal, whereas the converse does not hold in general; see Examples 2.5 and 2.10. The above argument shows more generally the following proposition.

Proposition 2.4. *For some $t \in \mathbb{N}$, let $F_{1,t}, \dots, F_{k,t}$ be forecasters with information sets $\mathcal{A}_{1,t}, \dots, \mathcal{A}_{k,t}$ in a prediction space for serial dependence. If $F_{1,t}$ is cross-ideal with respect to $\mathcal{A}_{1,t}, \dots, \mathcal{A}_{k,t}$, then it is also cross-ideal with respect to $\mathcal{A}_{1,t}, \mathcal{A}_{i_2,t}, \dots, \mathcal{A}_{i_m,t}$, where $\{i_2, \dots, i_m\} \subset \{2, \dots, k\}$.*

For clarity, we have chosen to illustrate the notions of cross-ideal forecasters (or cross-calibrated forecasters; see Definition 2.7) with independent forecast-observation tuples, or, in other words, in the one-period prediction space setting of Gneiting and Ranjan (2013) dropping the time index t. This is natural, as the notions of calibration are essentially one-period concepts.

Example 2.5. Let ν be uniformly distributed on $(5, 20)$ and, conditionally on ν, let δ have an inverse chi-squared distribution with ν degrees of freedom. Conditional on ν and δ, the outcome Y is normally distributed with mean zero and variance δ, and we consider two forecasters, a normally distributed forecaster $F_1 = \mathcal{N}(0, \delta)$ and a t-distributed forecaster $F_2 = t_\nu$. This example is constructed such that F_1 has the full information about the distribution of the outcome Y, whereas F_2 only knows the prior distribution of δ. We have that F_1 and F_2 are both ideal with respect to their information sets $\mathcal{A}_1 = \sigma(\delta)$ and $\mathcal{A}_2 = \sigma(\nu)$, respectively, but only F_1 is cross-ideal with respect to $\mathcal{A}_1, \mathcal{A}_2$.

More specifically, the predictive density function $f_1(\cdot|\delta)$ of F_1 is a normal density with variance δ, and the predictive density function f_2 of F_2 is

$$f_2(x|\nu) = \int_0^\infty f_1(x|s) g(s|\nu)\, ds = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma(\nu/2)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \tag{2}$$

where $g(s|\nu) = (\nu/2)^{\nu/2} s^{\nu/2-1} \exp\{-\nu/(2s)\}/\Gamma(\nu/2)$ is the density function of an inverse chi-squared distribution with ν degrees of freedom. The right hand side of (2) is the density of a t-distribution. Equation (2) holds because for a normal likelihood with known mean, the inverse chi-squared distribution is a conjugate prior of a t-distributed posterior distribution. Therefore, we see that

F_1 is cross-ideal with respect to $\mathcal{A}_1, \mathcal{A}_2$. It is clear that F_2 is not cross-ideal with respect to $\mathcal{A}_1, \mathcal{A}_2$. We will come back to this example throughout the paper.

From now on, we often identify the Markov kernels $F_{i,t}$ with random cumulative distribution functions (CDF). More precisely, we write $F_{i,t}(\omega, x) = F_{i,t}(\omega, (-\infty, x])$ for all $x \in \mathbb{R}$. Often we will omit ω and write $F_{i,t}(x) = F_{i,t}(\omega, x)$. The σ-algebra generated by $F_{i,t}$, denoted by $\sigma(F_{i,t})$, is the smallest σ-algebra such that $\omega \mapsto F_{i,t}(\omega, x)$ is measurable for all $x \in \mathbb{Q}$.

Definition 2.6 (PIT)**.** Let F be a (possibly random) CDF, X be a random variable and V a standard uniform random variable independent of F and X. We define

$$Z_F^X = F(X-) + V\{F(X) - F(X-)\},$$

where $F(y-) = \lim_{x \uparrow y} F(x)$. In the prediction space for serial dependence, the random variable $Z_{i,t} := Z_{F_{i,t}}^{Y_{t+1}}$ is called the *probability integral transform (PIT)* of the i-th forecaster $F_{i,t}$.

The PIT is the most prominent diagnostic tool for checking calibration empirically (Dawid, 1984; Diebold et al., 1998). The random variable $Z_{i,t}$ takes values in $[0, 1]$. If F is deterministic and $X \sim F$, then Z_F^X is uniformly distributed and $F^{-1}(Z_F^X) = X$ almost surely, where F^{-1} is the quantile function of F; see for example Rüschendorf (2009). Based on the PIT we introduce the following notions of cross-calibration.

Definition 2.7 (cross-calibration)**.** Let $F_{1,t}, \ldots, F_{k,t}$ be forecasters in a prediction space for serial dependence. Let $\{i_1, \ldots, i_m\} \subset \{1, \ldots, k\}$.

1. The forecast $F_{1,t}$ is *cross-calibrated* with respect to $F_{i_1,t}, \ldots, F_{i_m,t}$ if

$$\mathcal{L}(Z_{1,t} | F_{i_1,t}, \ldots, F_{i_m,t}, \mathcal{T}_t) = \mathcal{U}([0,1]), \quad \text{almost surely,}$$

   where $\mathcal{U}([0, 1])$ denotes the uniform distribution on $[0, 1]$. To be precise, the left hand side is a Markov kernel $\kappa : \Omega \times \mathcal{B}(\mathbb{R}) \to [0, 1]$, which is required to be constant in $\omega \in \Omega$ and equal to the Lebesgue measure on $[0, 1]$.
2. For $1 \leq j \leq k$, $F_{1,t}$ is *marginally cross-calibrated* with respect to $F_{j,t}$ if

$$\mathbb{E}_{\mathbb{Q}} F_{j,t}(y) = \mathbb{Q}\{F_{j,t}^{-1}(Z_{1,t}) \leq y\},$$

   for all $y \in \mathbb{R}$.

For brevity, we sometimes speak of cross-calibration with respect to $\{i_1, \ldots, i_m\}$ instead of $F_{i_1,t}, \ldots, F_{i_m,t}$. Our definitions are natural generalizations of the notions of calibration for individual forecasters in Gneiting and Ranjan (2013, Definition 2.6), which we recall here for ease of comparison.

Definition 2.8 (calibration)**.** Let F be a forecaster for an outcome Y in a one-period prediction space.

1. The forecast F is *probabilistically calibrated* if Z_F^Y is uniformly distributed on $[0, 1]$.

2. The forecast F is *marginally calibrated* if $\mathbb{E}_{\mathbb{Q}} F(y) = \mathbb{Q}(Y \leq y)$ for all $y \in \mathbb{R}$.

In part 2 of Definition 2.8 the left-hand side of the equation depends only on the distribution of the forecast, whereas the right-hand side depends only on the distribution of the observation. Marginal calibration therefore assesses whether the average forecast distribution is equal to the marginal distribution of Y. If $F_{1,t}$ is marginally cross-calibrated with respect to $F_{j,t}$, then, on average, the PIT $Z_{1,t}$ of $F_{1,t}$ behaves like a standard uniform random variable when considered in view of $F_{j,t}$. Intuitively, this means that $F_{1,t}$ has enough information about $F_{j,t}$ and the observation Y_{t+1} to disguise itself as uniform on average when viewed through the eyes of $F_{j,t}$. Cross-calibration means that the PIT $Z_{1,t}$ of $F_{1,t}$ is uniformly distributed no matter what the other forecasters predict. In contrast, probabilistic calibration of F_1 means that Z_{F_1} is uniformly distributed on average over all possible predictions of the other forecasters, which is a weaker notion. The following theorem formally connects Definitions 2.7 and 2.8 showing that the former is indeed a generalization of the latter.

Theorem 2.9. *Consider forecasters $F_{1,t}, \ldots, F_{k,t}$ in a prediction space for serial dependence.*

1. *The forecast $F_{1,t}$ is marginally cross-calibrated with respect to itself, if and only if $F_{1,t}$ is marginally calibrated.*
2. *If $F_{1,t}$ is cross-calibrated with respect to $F_{i_1,t}, \ldots, F_{i_m,t}$, then $F_{1,t}$ is cross-calibrated with respect to any subset of $\{i_1, \ldots, i_m\}$. In particular, $F_{1,t}$ is cross-calibrated with respect to the empty set, that is, probabilistically calibrated.*
3. *If $F_{1,t}$ is cross-calibrated with respect to $F_{2,t}$, then it is also marginally cross-calibrated with respect to $F_{2,t}$.*

It is possible that a forecaster is marginally calibrated but not probabilistically calibrated; see Gneiting and Ranjan (2013, Example 2.4) which we take up below in Example 2.10 to illustrate cross-calibration. In contrast, the last claim of Theorem 2.9 shows that marginal cross-calibration with respect to a different forecaster is a necessary condition for cross-calibration.

Example 2.10. Let μ be standard normally distributed, which we denote by $\mu \sim \mathcal{N}(0, 1)$. Conditional on μ, the outcome is $Y \sim \mathcal{N}(\mu, 1)$. Let τ take the values 1 or -1 with equal probability, independent of Y and μ. We consider four forecasters $F_1, \ldots, F_4$ of different skill, whose properties are summarized in Table 1.

It is clear that the perfect forecaster F_1 is cross-calibrated with respect to F_1, F_2, F_3, F_4. It is straight forward to check that the climatological forecaster F_2 is not cross-calibrated with respect to any of F_1, F_3, F_4 but with respect to itself. As F_2 is deterministic, this corresponds to the fact that F_2 is ideal with respect to the trivial σ-algebra. As the sign-reversed forecaster F_4 is not probabilistically calibrated it cannot be cross-calibrated. The cross-calibration of F_3 with respect to F_1, F_2, F_4 is shown in Appendix C. The state-

TABLE 1
Properties of the forecasters of Gneiting and Ranjan (2013, Example 2.4)

Forecaster	Predictive distribution	Information set
Perfect	$F_1 = \mathcal{N}(\mu, 1)$	$\mathcal{A}_1 = \sigma(\mu)$
Climatological	$F_2 = \mathcal{N}(0, 2)$	$\mathcal{A}_2 = \{\emptyset, \Omega\}$
Unfocused	$F_3 = \frac{1}{2}\{\mathcal{N}(\mu, 1) + \mathcal{N}(\mu + \tau, 1)\}$	$\mathcal{A}_3 = \sigma(\mu, \tau)$
Sign-reversed	$F_4 = \mathcal{N}(-\mu, 1)$	$\mathcal{A}_4 = \sigma(\mu)$

Forecaster	Cross-calibration	Marginal cross-calibration wrt			
		F_1	F_2	F_3	F_4
Perfect	wrt F_1, F_2, F_3, F_4	yes	yes	yes	yes
Climatological	wrt F_2	no	yes	no	no
Unfocused	wrt F_1, F_2, F_4	yes	yes	no	yes
Sign-reversed	no	no	no	no	yes

Note: Further details are given in Example 2.10. Cross-calibration with respect to (wrt) F_2 is equivalent to cross-calibration with respect to $\mathcal{A}_2 = \{\emptyset, \Omega\}$, that is, probabilistic calibration.

ments about marginal cross-calibration in Table 1 are consequences of Theorem 2.9.

Example 2.10 has originally been constructed by Gneiting et al. (2007) (see also Hamill, 2001) to illustrate the limitations of assessing uniformity of PIT histograms, that is, probabilistic calibration, in the quest for ideal forecasters. By Theorem 2.9, the forecaster F_3 is in particular probabilistically calibrated but it cannot be ideal as it is not marginally calibrated. While the example is constructed in the framework of a one-period prediction space, its message can also be transferred to a time series context. It is important to note that the same limitations apply to the concept of a cross-calibrated forecaster if the forecaster itself is not contained in the conditioning set. We give further details on these points below, after Proposition 2.11.

Tsyplakov (2011, 2013) introduced a slightly more restrictive notion than an ideal forecaster, which is an *auto-calibrated* forecaster, that is, it fulfils $\mathcal{L}(Y \mid F) = F$, almost surely, in the one-period prediction space setting. Generally, an auto-calibrated forecaster is ideal with respect to $\sigma(F)$. Gneiting and Ranjan (2013) contend that it is unlikely that empirical test of auto-calibration are feasible, except for very special circumstances such as forecasts for binary random variables. In cases where forecasters are restricted to specific classes of distributions Held et al. (2010) have taken on the challenge to derive statistical tests for ideal forecasters in the sense of auto-calibration based on a score regression approach; for earlier work in this direction see Hamill (2001); Mason et al. (2007). In Section 3 of the supplementary material, we show that it is possible to extend the score regression approach of Held et al. (2010) to test for cross-ideal forecasters with respect to $\sigma(F_1), \ldots, \sigma(F_k)$.

In this paper, we challenge the statement of Gneiting and Ranjan (2013) by proposing two powerful tests for cross-calibration under very general assumptions that are justified even under serial dependence; see Sections 6.1 and 6.2. Note that the following Proposition 2.11 shows that auto-calibration is in fact a special case of cross-calibration.

Proposition 2.11. *Consider forecasters $F_{1,t}, \dots, F_{k,t}$ in a prediction space for serial dependence. Let $\{i_1, \dots, i_m\} \subset \{1, \dots, k\}$. Then, the following are equivalent:*

1. *The forecaster $F_{1,t}$ is cross-calibrated with respect to $F_{i_1,t}, \dots, F_{i_m,t}$.*
2. *For all $z \in [0,1)$, conditional on $F_{i_1,t}, \dots, F_{i_m,t}, \mathcal{T}_t$, the random variable $\mathbb{1}(Z_{1,t} \leq z)$ is Bernoulli distributed with parameter z.*

If $1 \in \{i_1, \dots, i_m\}$, then part one and two are equivalent to $F_{1,t}$ being cross-ideal with respect to $\sigma(F_{i_1,t}), \dots, \sigma(F_{i_m,t})$.

In a time series context where forecasters solely base their predictions on past realizations of the quantity of interest Y_t, hence their information set is $\mathcal{T}_t$ and $\sigma(F_{i,t}) \subset \mathcal{T}_t$, Proposition 2.11 shows that probabilistic calibration is equivalent to auto-calibration. If forecasters have larger information sets, such as some of the forecasters in Example 2.10, cross-calibration of $F_{1,t}$ is equivalent to being cross-ideal if the forecaster itself is contained in the information set. If this condition is not fulfilled, we do not see a clear interpretation of cross-calibration. As in the case of probabilistic calibration, this is illustrated by the unfocused forecaster F_3 in Example 2.10. The forecaster F_3 is probabilistically calibrated and even cross-calibrated with respect to F_1, F_2 and F_4 but we know from its construction that it is not a particularly good forecast, so it is not clear what cross-calibration really says about F_3 in this context.

While we do not recommend to assess cross-calibration for forecast evaluation if the forecaster itself is not contained in the information set, we allow for this possibility in Definition 2.7 for sake of generality, and to be able to clarify the possible pitfalls of a seemingly natural concept.

The following Lemma 2.12 and the unfocused forecaster F_3 of Example 2.10 show that the condition $1 \in \{i_1, \dots, i_m\}$ for the equivalence between cross-calibration and being cross-ideal is essential. The forecaster F_3 is cross-calibrated with respect to the other forecasters but not with respect to herself, and thus, cannot be cross-ideal with respect to the other forecasters or herself by the lemma.

Lemma 2.12. *Consider forecasters $F_{1,t}, F_{2,t}$ in a prediction space for serial dependence. If $F_{1,t}$ is cross-ideal with respect to $\sigma(F_{2,t})$, then it is also cross-ideal with respect to $\sigma(F_{1,t}, F_{2,t})$.*

We conclude this section with the announced generalization of the result of Diebold et al. (1998) on uniformity and independence of PIT values in a prediction space for serial dependence. To this end, we consider the following assumption which may be called an *independent information condition*.

Assumption 2.13. In a prediction space for serial dependence, assume that, for all $t \in \mathbb{N}$, $m \geq 1$,

$$\mathcal{L}(Y_{t+1} \mid \mathcal{A}_{1,t+m}, \dots, \mathcal{A}_{k,t+m}, \mathcal{T}_t) = \mathcal{L}(Y_{t+1} \mid \mathcal{A}_{1,t}, \dots, \mathcal{A}_{k,t}, \mathcal{T}_t). \tag{3}$$

Assumption 2.13 formalizes the idea that information from other sources about the outcome at time point $t + 1 + m$ should not influence the outcome

Y_{t+1} at time point $t+1$. Let us illustrate this point in the context of weather forecasting. Suppose a numerical weather prediction system is used to calculate the state of the atmosphere to help us predict temperature tomorrow. Such forecasts rely on many other variables than just the past temperatures, which are encoded in the σ-algebras $\mathcal{A}_{i,t}$. Assumption 2.13 means that if we let the numerical system run longer to give us also information about the atmosphere the day after tomorrow, this will have no influence on what temperature is realized tomorrow.

A different scenario where forecasters use additional information other than past realizations of the quantity of interest are for example the inflation rate predictions issued by the Bank of England (BoE); see Section 7 for details. In this case, parameters of the predictive two-piece normal distributions are decided by the BoE's Monetary Policy Committee based on estimation and expert opinion. The expert opinion is contained in the σ-algebras $\mathcal{A}_{i,t}$.

Theorem 2.14. *Suppose we are in the prediction space setting for serial dependence and Assumption 2.13 holds. Let* $\{i_1, \dots, i_m\} \subset \{1, \dots, k\}$ *and assume that* $F_{1,t} = \mathcal{L}(Y_{t+1} | \mathcal{A}_{i_1,t}, \dots, \mathcal{A}_{i_m,t}, \mathcal{T}_t)$ *for all* $t \in \mathbb{N}$. *Then, for all* $l \in \mathbb{N}_0$, *we have*

$$\mathcal{L}(Z_{1,t}, \dots, Z_{1,t+l} | \mathcal{A}_{i_1,t+l}, \dots, \mathcal{A}_{i_m,t+l}) = \mathcal{U}([0,1])^{\otimes(l+1)}, \quad \textit{almost surely,}$$

for all $t \in \mathbb{N}$. *Here,* $\mathcal{U}([0,1])^{\otimes(l+1)}$ *denotes the distribution of* $l+1$ *independent standard uniform random variables.*

Remark 1. If we consider q-step ahead forecasts for some $q \geq 2$, then the above result continues to hold for all vectors of the form

$$(Z_{1,t}, Z_{1,t+q}, \dots, Z_{1,t+mq}).$$

However, there may be dependence amongst $(Z_{1,t}, Z_{1,t+1}, \dots, Z_{1,t+q-1})$, which complicates matters when testing for cross-calibration. This problem also arises in tests for uniformity and independence of PIT values as suggested by Diebold et al. (1998). Several approaches to deal with this issue have been suggested in the literature; see Knüppel (2015) and references therein. In this paper, we restrict our attention to cross-calibration of one-period ahead forecasts but extensions to q-step ahead forecasts would certainly be of great interest.

3. Scoring rules, calibration and sharpness

In this section, we comment on the relation of proper scoring rules to the notions of cross-calibration. A scoring rule is a real-valued function $S(F, y)$ that takes a CDF F as the first and a real number y as the second argument; see e.g. Gneiting and Raftery (2007). The expected score under a CDF G is then $\mathbf{S}(F, G) = \int S(F, y)\, \mathrm{d}G(y)$. A scoring rule is called proper if the divergence $\mathbf{d}(F, G) = \mathbf{S}(F, G) - \mathbf{S}(G, G)$ is non-negative and strictly proper if $\mathbf{d}(F, G) = 0$ implies $G = F$. For simplicity, we consider a one-period prediction space with forecasters

F_1 and F_2, and we define $G = \mathcal{L}(Y)$, $G_1 = \mathcal{L}(Y|F_1)$, $G_2 = \mathcal{L}(Y|F_2)$. Bröcker (2009) shows the following decomposition for a strictly proper scoring rule S,

$$\mathbb{E}_{\mathbb{Q}} S(F_j, Y) = \mathbf{S}(G, G) - \mathbb{E}_{\mathbb{Q}} \mathbf{d}(G, G_j) + \mathbb{E}_{\mathbb{Q}} \mathbf{d}(F_j, G_j), \quad \text{for } j = 1, 2. \tag{4}$$

While Bröcker (2009) assumes a finite outcome space, the result should readily generalize to outcomes on $\mathbb{R}$ if S is sufficiently regular in the second argument. The term $\mathbf{S}(G, G)$ depends on the outcome only and may be interpreted as the uncertainty of Y. It is also called the entropy of the distribution G. The last term on the right hand side of (4), the expected divergence between the forecast F_j and the conditional distribution of Y on F_j, is non-negative and zero, if and only if, F_j is cross-calibrated with respect to F_j, and is called *reliability* term. The second term of the right hand side of (4) enters negatively. It describes the expected deviation between the distribution of Y and the conditional distribution of Y on F_j. It quantifies how sharp the forecaster F_j is in comparison to a climatological forecaster, predicting $G = \mathcal{L}(Y)$, and is called *resolution* term.

If the forecasters F_1 and F_2 are cross-calibrated with respect to themselves, that is, auto-calibrated, then the expected score difference, $\mathbb{E}_{\mathbb{Q}} S(F_2, Y) - \mathbb{E}_{\mathbb{Q}} S(F_1, Y)$ reflects their difference in sharpness, which may be interpreted as a justification of the principle to "maximize sharpness subject to calibration" formulated by Murphy and Winkler (1987); Gneiting and Raftery (2007). To the best of our knowledge, it is currently not well understood how a particular scoring rule orders forecasts with respect to sharpness. Therefore, even for auto-calibrated forecasts, different scoring rules may yield different forecast rankings.

If F_1 is cross-calibrated with respect to F_1, F_2, then one can show that

$$\begin{aligned} \mathbb{E}_{\mathbb{Q}} S(F_2, Y) - \mathbb{E}_{\mathbb{Q}} S(F_1, Y) &= -\mathbb{E}_{\mathbb{Q}} \mathbf{d}(G, G_2) + \mathbb{E}_{\mathbb{Q}} \mathbf{d}(G, F_1) + \mathbb{E}_{\mathbb{Q}} \mathbf{d}(F_2, G_2) \\ &= \mathbb{E}_{\mathbb{Q}} \mathbf{S}(G_2, G_2) - \mathbb{E}_{\mathbb{Q}} \mathbf{S}(F_1, F_1) + \mathbb{E}_{\mathbb{Q}} \mathbf{d}(F_2, G_2) \\ &= \mathbb{E}_{\mathbb{Q}} \mathbf{S}(G_2, F_1) - \mathbb{E}_{\mathbb{Q}} \mathbf{S}(F_1, F_1) + \mathbb{E}_{\mathbb{Q}} \mathbf{d}(F_2, G_2) \\ &= \mathbb{E}_{\mathbb{Q}} \mathbf{d}(G_2, F_1) + \mathbb{E}_{Q} \mathbf{d}(F_2, G_2). \end{aligned} \tag{5}$$

In particular, F_1 is preferred over F_2 by any strictly proper scoring rule S. This follows also from Holzmann and Eulert (2014, Theorem 3). However, the score decomposition approach allows to interpret the components leading to a difference in expected score. The first term on the right hand side of (5) is a penalty for the auto-calibrated forecast G_2 because it uses a smaller information set than F_1. The second term penalizes a potential lack of auto-calibration of F_2. The decomposition shows that in the presence of a cross-calibrated forecaster with respect to all forecasters, sharpness does not need to be considered as a further criterion for forecast selection as the cross-calibrated forecaster is automatically the sharpest with respect to any proper scoring rule.

Cross-calibration of F_1 with respect to F_2 is a weaker requirement than F_1 being cross-ideal with respect to $\sigma(F_1, F_2)$, which is in turn weaker than being cross-ideal with respect to $\sigma(F_2)$ by Lemma 2.12; see also Theorem 2.9 and Example 2.10. This is analogous to the fact that probabilistic calibration of F_1 (cross-calibration with respect to $\{\emptyset, \Omega\}$) is a weaker notion than auto-

calibration (cross-calibration with respect to F_1). Unfortunately, we are currently not able to describe the effect of probabilistic calibration of F_1 or of cross-calibration of F_1 with respect to F_2 on the expected score in general, other than being necessary requirements for auto-calibration and being cross-ideal with respect to $\sigma(F_1, F_2)$, respectively. The main merit of these necessary requirements lies in the fact that they can be assessed empirically.

4. Binary outcomes

In this section, we consider the case when the observation Y only takes two values, zero and one. We interpret $Y = 1$ as a success and $Y = 0$ as a failure. A forecaster F is then represented by her predictive success probability p, such that the predictive CDF is $F(y) = p \cdot \mathbb{1}(y \geq 1) + (1 - p) \cdot \mathbb{1}(y \geq 0)$. We identify F with p, where p is a random variable taking values in $[0, 1]$.

In the case of an individual forecaster F for a binary outcome, it has been shown in Gneiting and Ranjan (2013, Theorem 2.11) that the notions of a probabilistically calibrated forecaster F and an ideal forecaster relative to the σ-algebra generated by the predictive probability p are equivalent. Furthermore, both notions coincide with the notion of *conditional calibration*, that is $\mathbb{Q}(Y = 1 \,|\, p) = p$. This result carries over to the notions of cross-calibration of multiple forecasters introduced in this paper. As the notions of calibration are essentially only concerned with one prediction period, we have chosen to present the results of this section in the one-period prediction space setting of Definition 2.1 for simplicity.

Theorem 4.1. *Consider the one-period prediction space setting with binary outcome Y and forecasts $F_1, \ldots, F_k$ represented by their predictive success probabilities $p_1, \ldots, p_k$, respectively. Then the following statements are equivalent:*

1. *The forecast p_1 is cross-calibrated with respect to $p_2, \ldots, p_k$, that is $\mathcal{L}(Z_{p_1} \,|\, p_2, \ldots, p_k)$ is standard uniform.*
2. *The forecast p_1 is conditionally cross-calibrated with respect to $p_1, \ldots, p_k$, that is $\mathbb{Q}(Y = 1 \,|\, p_1, \ldots, p_k) = p_1$.*
3. *The forecast p_1 is ideal relative to $\sigma(p_1, \ldots, p_k)$.*

The cross-calibration notion of Feinberg and Stewart (2008) is analogous to our notion of cross-calibration with respect to $\{1, \ldots, k\}$ which is equivalent to cross-ideal forecasters for binary events. Theorem 4.1 shows that both notions coincide with cross-calibration of p_1 with respect to $\{2, \ldots, k\}$ which is a priori a weaker requirement; see also Lemma 2.12 and Example 2.10. As noted by Gneiting and Ranjan (2013) the fact that probabilistically calibrated forecasters are automatically auto-calibrated clarifies the relation between PIT-histograms and *calibration curves* which are the diagnostic tool frequently used for assessing calibration of binary predictions (Dawid, 1986; Murphy and Winkler, 1992; Ranjan and Gneiting, 2010). As described in Section 5, cross-calibration can be assessed with conditional PIT-histograms. Analogously, in the case of binary forecasts, conditional calibration curves can be considered.

5. Diagnostic plots for assessing cross-calibration

Gneiting et al. (2007) suggest to assess marginal calibration based on a plot of the empirical analogue of the difference

$$\mathbb{E}_{\mathbb{Q}} F_t(y) - \mathbb{Q}(Y_{t+1} \le y), \quad \text{for } y \in \mathbb{R}.$$

Analogously, to assess marginal cross-calibration, the empirical version of

$$\mathbb{E}_{\mathbb{Q}} F_{j,t}(y) - \mathbb{E}_{\mathbb{Q}} \mathbb{1}\{F_{j,t}^{-1}(Z_{i,t}) \le y\}, \quad \text{for } y \in \mathbb{R}, \tag{6}$$

can be plotted. If the graph is significantly different from a horizontal line through zero, one can deduce that $F_{i,t}$ is not marginally cross-calibrated with respect to $F_{j,t}$ and therefore also not cross-calibrated with respect to $F_{j,t}$ by Theorem 2.9. If the graph is not significantly different from zero anywhere, then we have marginal cross-calibration. However, this does not necessarily imply that we have a cross-calibrated forecaster.

Probabilistic calibration is often checked empirically by plotting a histogram of $Z_{i,t}$, the so-called PIT-histogram. Generally, it is not obvious how to check cross-calibration empirically. However, in many situations of practical interest it can be done by borrowing the idea of considering forecasting profiles as in the cross-calibration test of Feinberg and Stewart (2008). Suppose that the forecasters $F_{1,t}, \dots, F_{k,t}$ pick predictions from some parametric class of distributions $\mathcal{F} = \{F_{\boldsymbol{\lambda}} \mid \boldsymbol{\lambda} \in \Lambda\}$, where $\Lambda \subset \mathbb{R}^d$. Then we can identify each forecaster $F_{i,t}$ with the parameter $\boldsymbol{\lambda}_{i,t}$ she predicts. We observe a sample $(F_{1,t}, \dots, F_{k,t}, Y_{t+1}, V_t)$ for $1 \le t \le N$. Let $\Lambda_1, \dots, \Lambda_p$ be a partition of the parameter space. For a diagnostic plot showing if $F_{1,t}$, say, is cross-calibrated with respect to $\{i_1, \dots, i_m\}$, we can sort the observations into pm bins according to the predicted values $\boldsymbol{\lambda}_{i_1,t}, \dots, \boldsymbol{\lambda}_{i_m,t}$. Then a PIT-histogram of $Z_{1,t}$ can be plotted for each bin. Clearly, the number of bins needs to be small in relation to the number of observations. Partitioning the parameter space links cross-calibration to the stratification approach of Murphy (1994).

We illustrate these diagnostic tools with two examples. Further illustrations are provided in the supplementary material.

Example 5.1 (Example 2.10 continued)**.** In Figure 1 the differences given at (6) are plotted for the four forecasters of Example 2.10. More precisely, the random variables are simulated 10'000 times and the empirical expectation is plotted. Recall that, for all simulation examples we are using independent forecast-observation tuples for reasons of simplicity. In this example, it is easy to see that F_1 is superior to F_2 using the notion of marginal cross-calibration, which was not the case using only the calibration notions of Gneiting and Ranjan (2013, Definition 2.6); see Gneiting et al. (2007).

Example 5.2 (Example 2.5 continued)**.** Coming back to the forecasters F_1 and F_2 of Example 2.5, PIT-histograms for assessing cross-calibration with respect to F_1, F_2 for 10'000 simulations are given in Figure 2. To show the lack of cross-calibration of F_2 with respect to F_1, F_2 it is sufficient to consider δ in order to

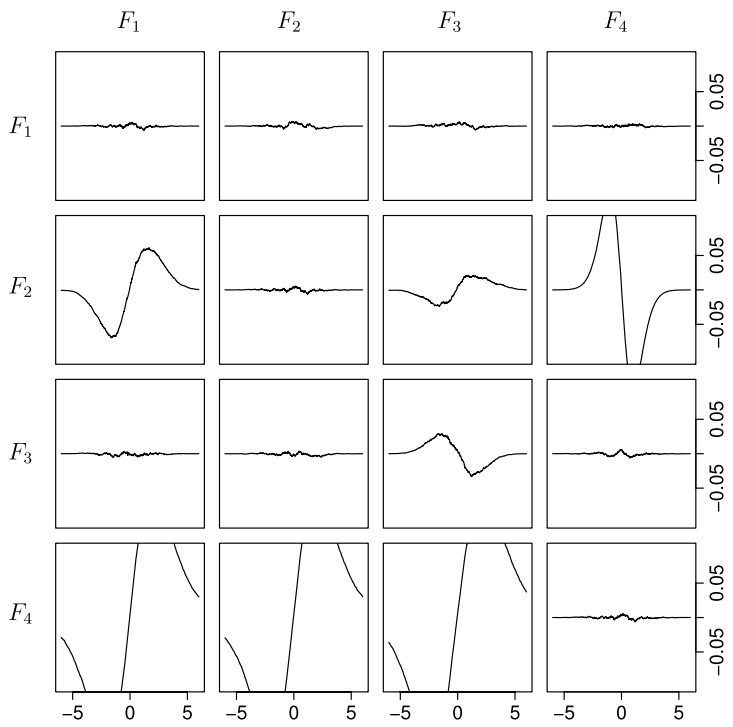

FIG 1. *Marginal cross-calibration plots of the forecasters in Example 2.10 with 10'000 simulations. In the i-th row and j-th column the empirical version of Equation (6) is plotted to assess whether F_i is marginally cross-calibrated with respect to F_j or not.*

choose the partition. The intervals are chosen such that in each histogram there are around the same amount of observations.

6. Tests for assessing cross-calibration

In this section, we consider statistical tests for cross-calibration. The tests in Section 6.1 are based on the idea of conditional exceedance probabilities (Mason et al., 2007), whereas the tests in Section 6.2 use a linear regression approach. Suppose we have observations $F_{1,t}, \dots, F_{k,t}$ and Y_{t+1}, $1 \leq t \leq N$ in a prediction space for serial dependence. We would like to test the null hypothesis that $F_{1,t}$ is cross-calibrated with respect to $J \subset \{1, \dots, k\}$. Under the null hypothesis, using Proposition 2.11, conditional on $F_{i,t}$ for all $i \in J$, the random variable $Z_{1,t}$ is standard uniformly distributed. Under the independent information condition, Assumption 2.13, the random variables $Z_{1,1}, \dots, Z_{1,N}$ are independent conditional on $\mathcal{A}_{i,N}$ for all $i \in J$ by Theorem 2.11. Clearly, they are also independent in the special case of independent forecast-observation-tuples $(F_{1,t}, \dots, F_{k,t}, Y_{t+1})$, $1 \leq t \leq N$. In this section, we assume that Assumption 2.13 holds.

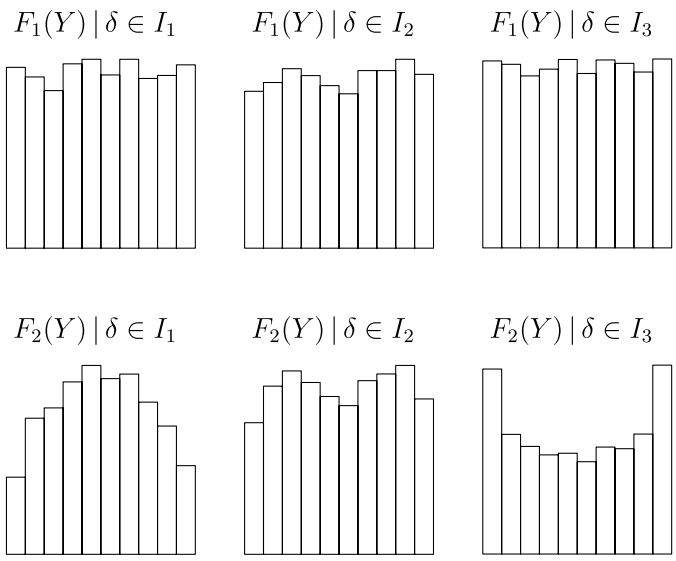

FIG 2. *PIT-histogram plots of F_1 in the top row and F_2 in the bottom row conditional on δ, respectively, where $I_1 = [0, 0.95], I_2 = (0.95, 1.1], I_3 = (1.1, \infty]$.*

The Tables in this section and Section 7 show the results of the proposed tests as Monte-Carlo powers and p-values, respectively. The numbers in square brackets represent tests for cross-calibration with respect to the trivial information set, that is probabilistic calibration, in the case that the forecaster under consideration is using a larger information set. The numbers in round brackets represent tests for cross-calibration with respect to sets of forecasters without the forecaster under consideration. (Strictly speaking, the first case is a special case of second. However, due to the widespread use of probabilistic calibration tests, we treat this case separately.) We have decided to put the results in brackets to recall our cautionary remarks of Section 2 on the interpretation of cross-calibration in these cases. Note that the formulation and implementation of the tests can be done in this generality without additional effort.

6.1. Conditional exceedance probabilities

For $z \in (0, 1)$, we define $B_{z,t} := \mathbb{1}\{Z_{1,t} \leq z\}$. We stipulate the logistic regression models

$$\operatorname{logit}[\mathbb{Q}\{B_{z,t} = 1 | F_{i,t}^{-1}(z), i \in J\}] = \beta_{0,z} + \sum_{i \in J} \beta_{i,z} F_{i,t}^{-1}(z) \tag{7}$$

for each $z \in (0, 1)$, where $\operatorname{logit}(p) = \log\{p/(1-p)\}$ is the logistic function. Using (7), the null hypothesis is

$$H_0 : \quad \beta_{0,z} = \operatorname{logit}(z),\ \beta_{i,z} = 0,\ i \in J, \quad \text{for all } z \in (0, 1). \tag{8}$$

Generally, we have that $\sigma(F_{i,t}) \subset \sigma(\mathcal{A}_{i,t}, \mathcal{T}_t)$. If $\sigma(F_{i,t}) \subset \mathcal{A}_{i,t}$ holds, at least approximately, it is reasonable to assume that the Bernoulli random variables $B_{z,1}, \ldots, B_{z,N}$, $1 \leq t \leq N$ in (7) are conditionally independent. In the case of independent forecast-observation-tuples this is clearly fulfilled. Alternatively, if the information in $\mathcal{A}_{i,t}$ is given separately from $\mathcal{T}_t$, then one can formulate (7) conditional on this information instead of conditioning on $F_{i,t}^{-1}(z)$, $i \in J$, to ensure conditional independence.

For each $z \in (0,1)$, we suggest to test the pointwise hypothesis

$$H_0(z): \quad \beta_{0,z} = \text{logit}(z),\ \beta_{i,z} = 0,\ i \in J, \tag{9}$$

by a likelihood ratio test yielding a p-value $\pi(z)$. More precisely, the covariate vector $\mathbf{x}_{z,t}$ has one as the first entry and then $F_{i,t}^{-1}(z)$, $i \in J$ and the parameter vector $\boldsymbol{\beta}_z$ has entries $\beta_{0,z}$, $\beta_{i,z}$, $i \in J$. For values of z close to zero or one, we frequently encounter the phenomenon of separation, that is, the likelihood converges, but at least one parameter value is infinite. Therefore, we have chosen to use the method of Firth (1993), which always yields finite parameter estimates; see Heinze and Schemper (2002). That is, we fit the parameters $\beta_{0,z}$, $\beta_{i,z}$, $i \in J$ by maximizing the penalized log-likelihood function

$$\ell_p(\boldsymbol{\beta}_z) = \ell(\boldsymbol{\beta}_z) + \frac{1}{2}\log|I(\boldsymbol{\beta}_z)|,$$

where

$$\ell(\boldsymbol{\beta}_z) = \sum_{t=1}^{N} B_{z,t}\mathbf{x}_{z,t}^\top \boldsymbol{\beta}_z - \sum_{t=1}^{N} \log\{1 + \exp(\mathbf{x}_{z,t}^\top \boldsymbol{\beta}_z)\},$$

and $|I(\boldsymbol{\beta}_z)|$ is the determinant of the Fisher information matrix. We denote the estimated parameter vector by $\hat{\boldsymbol{\beta}}_z$ with entries $\hat{\beta}_{0,z}$, $\hat{\beta}_{i,z}$, $i \in J$. For N large enough, the test statistic

$$T_z = -2\{\ell_p(\hat{\boldsymbol{\beta}}_z) - \ell_p(\boldsymbol{\gamma}_z)\}$$

has a χ^2-distribution with $1 + |J|$ degrees of freedom, where $|J|$ denotes the cardinality of $|J|$, and $\boldsymbol{\gamma}_z = (\text{logit}(z), 0, \ldots, 0)^\top$. We define the p-value $\pi(z) = 1 - \chi^2_{1+|J|}(T_z)$, where $\chi^2_{1+|J|}$ denotes the cumulative distribution function of a χ^2 random variable with $1 + |J|$ degrees of freedom. For the simulation studies below and the data analysis in Section 7 we have used the R-package of Heinze et al. (2013) to calculate T_z.

In order to draw conclusions about the global null hypothesis H_0 at (8) from the pointwise p-values $\pi(z)$, we adjust them for multiple testing. We follow the approach of Cox and Lee (2008) to use the method of Westfall and Young (1993, Chapter 2) for functional data to compute adjusted p-values $r(z)$; see also Meinshausen et al. (2011).

Let $0 < z_1 < \cdots < z_M < 1$. Under the null hypothesis of cross-calibration, it is possible to simulate a vector of p-values $(\pi^*(z_1), \ldots, \pi^*(z_M))$ with the same distribution as $(\pi(z_1), \ldots, \pi(z_M))$ conditional on $F_{i,t}^{-1}(z_m)$, $i \in J$, $1 \leq t \leq N$, $1 \leq m \leq M$, as follows. Let $U_1, \ldots, U_N$ be iid standard uniform random

variables. For $1 \le m \le M$, define $B^*_{z_m,t} = \mathbb{1}(U_t \le z_m)$, and let $\pi^*(z_m)$ be the p-value from the pointwise likelihood ratio test for the simulated data vector $(B^*_{z_m,t})_{1\le t\le N}$ and covariates $(\mathbf{x}_{z_m,t})_{1\le t\le N}$ as before.

The adjusted p-values can now be obtained as follows. Let ρ be the permutation of $\{1,\dots,M\}$ such that $\pi(z_{\rho(1)}) \le \cdots \le \pi(z_{\rho(M)})$. This permutation ρ remains unchanged in the following procedure. For a simulated vector of p-values $(\pi^*(z_1),\dots,\pi^*(z_M))$, we define $q^*_m = \min\{\pi^*(z_{\rho(s)})\colon s \ge m\}$. Repeating this procedure L times, we obtain an array $(q^*_{m,l})_{1\le m\le M,1\le l\le L}$ and define the adjusted p-values $r_1,\dots,r_M$ corresponding to $z_1,\dots,z_M$ as

$$r_m = \frac{1}{L}\sum_{l=1}^{L} \mathbb{1}\{q^*_{\rho^{-1}(m),l} \le \pi(z_m)\}, \quad 1 \le m \le M.$$

The global null hypothesis H_0 at (8) can be rejected at level $\alpha \in (0,1)$ if $\min[r_m\colon 1 \le m \le M] \le \alpha$. Furthermore, the adjusted p-values allow to draw conclusions for which values of $z_m \in (0,1)$ miscalibration occurs. For example, a prediction method may perform satisfactory for the left tail of the distribution, that is, for z close to zero, the adjusted p-values are large, whereas it fails to capture the right tail and hence for z close to one, the adjusted p-values are small. We call this test the *CEP test with respect to J*.

Remark 2. It is important to note that the adjusted p-values r_m remain the same, if the pointwise p-values $\pi(z)$ are transformed with a strictly monotone transformation. Therefore, even if the $\pi(z)$ are only asymptotic p-values, the adjusted p-values r_m will control the familywise error rate at the desired level α even for finite samples (for large numbers L of bootstrap replications); see Westfall and Young (1993, Chapter 2) and Cox and Lee (2008). It is nevertheless important which test statistic to choose for the pointwise tests as the power of the overall test will crucially depend on the power of the pointwise tests.

Example 6.1 (Example 2.10 continued)**.** We consider the forecasters $F_1,\dots,F_4$ of Example 2.10; see Table 1. For sample size $N = 50$, we performed the CEP tests for $F_1,\dots,F_4$ with respect to all possible subsets of $F_1,\dots,F_4$ at significance level $\alpha = 0.05$ and calculated the Monte Carlo power based on $10'000$ simulations. We used the gridpoints $z_m = \{1+(18/19)m\}/20$, $0 \le m \le 19$. The number of bootstrap replications for calculating the adjusted p-values is set to $L = 500$. For data examples, L should be much larger. However, for analyzing the performance of the resampling based p-values, it is more important to run a large number of simulations than to have a large bootstrap sample for each of them; see Westfall and Young (1993) for a more detailed discussion. The results are given in Table 2.

Conditioning on F_2 corresponds to conditioning on the trivial σ-algebra, therefore testing conditional on F_1, F_2, F_3 is the same as testing conditional on F_1, F_3, for example. Hence, Table 2 contains all interesting subsets of $F_1,\dots,F_4$ and the column entitled 'F_2' corresponds to a test for probabilistic calibration. The test performs well, even for the small sample size $N = 50$. Generally, the power of the test appears to increase, the more information is used. For exam-

TABLE 2
Monte Carlo power for the CEP tests in Example 6.1

wrt	F_1	F_2	F_3	F_4	F_1, F_3	F_1, F_4	F_3, F_4	F_1, F_3, F_4
F_1	0.056	[0.051]	(0.055)	(0.055)	0.050	0.051	(0.050)	0.048
F_2	(0.997)	0.051	(0.979)	(0.997)	0.994	0.990	0.993	0.977
F_3	(0.052)	[0.052]	0.168	(0.051)	0.635	(0.051)	0.634	0.582
F_4	(1.000)	[1.000]	(1.000)	1.000	(1.000)	1.000	1.000	1.000

TABLE 3
Monte Carlo power for the CEP tests in Example 6.2.

	$N = 50$			$N = 200$		
wrt	F_1	F_2	F_1, F_2	F_1	F_2	F_1, F_2
F_1	0.052	(0.053)	0.050	0.050	(0.052)	0.051
F_2	(0.156)	0.051	0.139	(0.533)	0.057	0.464

ple, the test has difficulty to detect that F_3 is not ideal with respect to itself but it performs well for rejecting the null hypothesis that F_3 is cross-ideal with respect to F_1, F_3, F_3, F_4 or F_1, F_3, F_4.

Prompted by the comments of a reviewer, we would like to issue a word of warning concerning the interpretation of the test results in this example. Suppose an applied researcher proceeds as follows: First, she tests for auto-calibration of F_3 and this is not rejected. Then she tests for cross-calibration of F_3 with respect to F_1, F_3 which is rejected. Note that this is a likely situation given the powers in Table 6.1. From a theoretical point of view it is intuitive that the power of testing with respect to F_1, F_3 is higher than when testing with respect to F_3 as the former null hypothesis is a subset of the latter. However, she could draw the somewhat misleading conclusion that F_3 is close to being auto-calibrated but that F_1 uses additional information that is not used in F_3. If she is interested in a true comparison of forecasters F_1 and F_3 she should have also tested for auto-calibration of F_1 and cross-calibration of F_1 with respect to F_1 and F_3. As both of these hypotheses cannot be rejected she should conclude correctly that F_1 uses the information contained in F_1 and F_3 better than F_3 and that this forecast should thus be preferred.

Example 6.2 (Example 2.5 continued)**.** We applied the CEP tests to data simulated from the prediction space described in Example 2.5 at significance level $\alpha = 0.05$. We used the same grid and other parameters as in the previous example, except that we considered two different sample sizes $N = 50$ and $N = 200$. The results from $10'000$ simulations can be seen in Table 3. Here, the power for sample size $N = 50$ is only small. Fortunately, it appears to increase rapidly with sample size and is satisfactory for $N = 200$.

6.2. Linear regression approach

To formulate the linear regression approach (LRA) tests for cross-calibration, we restrict ourselves to a parametric class of cumulative distribution functions

$\mathcal{F} = \{F_{\boldsymbol{\lambda}} | \, \boldsymbol{\lambda} \in \Lambda\}$, where $\Lambda \subset \mathbb{R}^d$. Each forecaster $F_{i,t}$ is then represented by the predictive parameter vector $\boldsymbol{\lambda}_{i,t} = (\lambda_{i,t}^{(1)}, \ldots, \lambda_{i,t}^{(d)})$ for $1 \le i \le k$. This leads to the null hypothesis

$$H_0: \quad \mathcal{L}\{\Phi^{-1}(Z_{1,1}), \ldots, \Phi^{-1}(Z_{1,N}) | \boldsymbol{\lambda}_{j,t}, j \in J, 1 \le t \le N\} = \mathcal{N}_N(0, I_N), \quad (10)$$

where Φ^{-1} denotes the quantile function of a standard normal distribution and $\mathcal{N}_N(0, I_N)$ denotes a multivariate standard normal distribution. Here, as for the CEP tests, some care has to be taken. In general, $\sigma(\boldsymbol{\lambda}_{i,t}) \subset \sigma(\mathcal{A}_{i,t}, \mathcal{T}_t)$. If $\sigma(\boldsymbol{\lambda}_{i,t}) \subset \mathcal{A}_{i,t}$ holds, at least approximately, it is reasonable to assume conditional independence of $Z_{1,1}, \ldots, Z_{1,N}$, $1 \le t \le N$ in (10). In other words, in (10), one should only condition on the parameters that are $\mathcal{A}_{i,t}$-measurable.

In order to test the hypothesis at (10) we perform an F-test based on linear regression. We consider the linear model

$$\mathbf{Y} = \mathbf{D_J}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad (11)$$

where $\mathbf{Y} = (\Phi^{-1}(Z_{1,1}), \ldots, \Phi^{-1}(Z_{1,N}))^T \in \mathbb{R}^N$ is the response vector,

$$\mathbf{D_J} = \begin{pmatrix} 1 & \lambda_{i_1,1}^{(1)} & \cdots & \lambda_{i_1,1}^{(d)} & \lambda_{i_2,1}^{(1)} & \cdots & \lambda_{i_m,1}^{(d)} \\ 1 & \lambda_{i_1,2}^{(1)} & \cdots & \lambda_{i_1,2}^{(d)} & \lambda_{i_2,2}^{(1)} & \cdots & \lambda_{i_m,2}^{(d)} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & \lambda_{i_1,N}^{(1)} & \cdots & \lambda_{i_1,N}^{(d)} & \lambda_{i_2,N}^{(1)} & \cdots & \lambda_{i_m,N}^{(d)} \end{pmatrix} \in \mathbb{R}^{N \times (1+dm)} \quad (12)$$

is the design matrix, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{dk})^T \in \mathbb{R}^{1+dm}$ is the parameter vector we would like to estimate, and $\boldsymbol{\epsilon} \in \mathbb{R}^N$ is a random error vector, which is multivariate standard normal under the null hypothesis. In order to estimate $\boldsymbol{\beta}$ the method of least squares is used and we obtain the estimated parameter vector $\hat{\boldsymbol{\beta}}$, the vector of fitted values $\hat{\boldsymbol{Y}} = \mathbf{D_J}\hat{\boldsymbol{\beta}}$, and the residual vector $\hat{\boldsymbol{\epsilon}} = \mathbf{Y} - \hat{\mathbf{Y}}$.

Under the null hypothesis we have that

$$\boldsymbol{\beta} = (0, 0, \ldots, 0)^T \in \mathbb{R}^{1+dm} \quad \text{and} \quad \boldsymbol{\epsilon} \sim \mathcal{N}_N(0, I_N).$$

To test the assumption that $\boldsymbol{\epsilon}$ is standard normal one can use a normality test such as the Anderson-Darling or Shapiro-Wilk (Anderson and Darling, 1954; Shapiro and Wilk, 1965; Yap and Sim, 2011). This yields a p-value π_{normal}. To test the other assumption we consider the test statistic

$$F_0 = \frac{\hat{\boldsymbol{\beta}}^T(\mathbf{D_J}^T\mathbf{D_J})\,\hat{\boldsymbol{\beta}}}{(1+dm)\hat{\sigma}^2},$$

where $\hat{\sigma}^2 = \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2/\{N - (1+dm)\}$ is the unbiased variance estimator. The test statistic F_0 has a Fisher distribution with $1 + dm$ and $N - 1 - dm$ degrees of freedom; see for example Montgomery et al. (2001). The p-value π_F is then $\pi_F = 1 - F_{1+dm,N-1-dm}(F_0)$, where $F_{p,q}$ denotes the Fisher cumulative

TABLE 4
Monte Carlo power of the LRA test in Example 6.3.

	$N = 20$			$N = 50$		
wrt	$\emptyset$	μ	μ, τ	$\emptyset$	μ	μ, τ
F_1	[0.026]	0.024	0.025	[0.022]	0.024	0.025
F_2	0.025	0.884	0.825	0.027	1.000	1.000
F_3	[0.025]	(0.024)	0.238	[0.023]	(0.026)	0.734
F_4	[0.880]	1.000	0.999	[1.000]	1.000	1.000

distribution function with p and q degrees of freedom. Combining these two tests by the method of Holm leads to the adjusted p-value

$$\pi_{\text{adjust}} = 2\min(\pi_F, \pi_{\text{normal}}).$$

We need that the rank of the design matrix $\mathbf{D_J}$ is $1 + dm$, otherwise the regression analysis is not possible. Therefore, any forecaster $F_{i,t}$, $i \in J$ has to predict at least two distinct values for each parameter. Otherwise, we omit the parameter for this forecaster in the model and are still able to use the test, which we call the *LRA test with respect to* J.

Example 6.3 (Example 2.10 continued)**.** Recall the forecasters F_1, F_2, F_3 and F_4 from Example 2.10. All four forecasters are in the class of distribution functions $\mathcal{F} = \{F_\lambda | \lambda = (\mu, \sigma, \tau) \in \mathbb{R} \times (0, \infty) \times \{-1, 0, 1\}\}$ for $F_\lambda = \frac{1}{2}\{\mathcal{N}(\mu, \sigma) + \mathcal{N}(\mu + \tau, \sigma)\}$. We apply the LRA test for sample sizes $N = 20$ and $N = 50$ at significance level $\alpha = 0.05$. The Monte Carlo powers of π_{adjust} for $10'000$ simulations are given in Table 4. In the LRA we condition on the parameters rather then the forecasts themselves. Therefore, $\emptyset$ corresponds to F_2, μ to F_1 and F_4 and μ, τ to F_3. Testing cross-calibration with respect to $J \subset \{1, 2, 3, 4\}$ leads then to the same test as testing with respect to $\emptyset$ if J is empty or $J = 2$, testing with respect to μ, τ if $3 \in J$ and testing with respect to μ otherwise. For testing standard normality, we have used an Andersen-Darling test (with mean set to zero and variance set to one). In the cases of cross-calibration, the normality test never rejects the null hypothesis, which explains the conservative levels of around 0.025 in these cases. The test is powerful even for the small sample sizes and it provides the expected results from the theoretical considerations; see Table 1. In particular, the LRA test detects well, that F_3 is not ideal with respect to itself contrary to the CEP test; compare Table 2.

Example 6.4 (Example 2.5 continued)**.** Coming back to forecasters F_1 and F_2 from Example 2.5 we perform the F-test for different sample sizes N. Here, δ corresponds to F_1 and ν corresponds to F_2. The Monte Carlo powers of the tests for $10'000$ simulations can be found in Table 5. The Monte Carlo powers are low even for large sample sizes, contrary to the results of the CEP tests; compare Table 3. We do not report the power of LRA test in this example because the Anderson-Darling test for standard normality almost never rejects the null hypothesis.

TABLE 5
Monte Carlo power for the F-test in the LRA in Example 6.4

	$N = 20$			$N = 50$		
wrt	δ	ν	δ, ν	δ	ν	δ, ν
F_1	0.051	(0.053)	0.050	0.049	(0.050)	0.048
F_2	(0.092)	0.051	0.081	(0.105)	0.050	0.092
	$N = 100$			$N = 200$		
wrt	δ	ν	δ, ν	δ	ν	δ, ν
F_1	0.048	(0.045)	0.046	0.049	(0.050)	0.049
F_2	(0.114)	0.045	0.097	(0.122)	0.048	0.108
	$N = 1'000$			$N = 5'000$		
wrt	δ	ν	δ, ν	δ	ν	δ, ν
F_1	0.053	(0.050)	0.051	0.047	(0.050)	0.048
F_2	(0.139)	0.051	0.121	(0.135)	0.049	0.119

6.3. Summary

We have presented two different approaches for testing cross-calibration, the CEP tests in Section 6.1 and the LRA tests in Section 6.2. Both approaches allow to test for cross-calibration of F_1 with respect to any subset $J \subset \{1, \dots, k\}$. The CEP test and the LRA test with respect to $J = \emptyset$ are tests for probabilistic calibration, that is, the classical hypothesis of uniformity and independence of PIT values. The tests are formulated in a prediction space for serial dependence, which is a scenario that is frequently encountered in practice; see also Section 7.

The CEP test has the advantage that it provides information concerning the parts of the distribution where miscalibration is detected (in terms of quantile levels); this is illustrated in Figures 3 and 4. It may be considered a disadvantage that the adjusted p-values are simulation based and depend on a grid $0 < z_1 < \cdots < z_M < 1$ that is to be chosen. In simulations, the method has shown to be robust to the number M of grid points. In contrast, the p-values for the LRA test are given explicitly. The forecasters have to be described through a finite-dimensional parameter vector and there are some restrictions concerning the predictive parameters, as it has to be ensured that the design matrix $\mathbf{D_J}$ at (12) has full rank. For the forecasters of Example 2.10, the LRA test has overall a better power than the CEP test; see Examples 6.1 and 6.3. The difference is minor, except for the hypothesis that the forecaster F_3 is ideal. Here, for sample size $N = 50$, the LRA test achieves a power of 0.734, whereas the CEP test only has a power of 0.168. For the forecasters in Example 2.5 the CEP test outperformed the LRA test; see Examples 6.2 and 6.4. In fact, for sample size $N = 200$, the power of the CEP test is more than three times higher than the power of the LRA test.

The following modifications of the CEP and the LRA tests are straight forward but unexplored. The logistic regression model in (7) can be replaced by any other regression model for a binary outcome variable, where it is possible to formulate a test for an analogous pointwise null hypothesis as given at (9). If forecasters choose their distributions from a parametric class of distributions

as assumed in LRA approach, it could also be considered to regress the random variables $B_z = \mathbb{1}(Z_{1,t} \leq z)$ on the predicted parameter values. In the LRA, the linear regression model stipulated at (11) can be replaced by some other regression model for a vector of real valued outcomes.

We would like to remark that the CEP and the LRA tests are formulated in the prediction space setting for serial dependence and make use of Assumption 2.13. It appears that deciding whether this assumption is justified in a given application context is sometimes a delicate matter. For example, if forecaster i bases her predictions purely on intuition, then Assumption 2.13 is certainly justified. If forecaster j uses a time series model for predictions, that is, predictions are exclusively derived from past data, then $\mathcal{A}_{j,t} = \{\emptyset, \Omega\}$ and Assumption 2.13 is trivially fulfilled. The CEP and LRA tests should only be applied with respect to sets J such that $j \notin J$. It may be that some parameters of a predictive distribution are derived from past data, whereas others are from external sources such as expert opinion. Here, one should only regress on the latter type of parameters in the LRA tests and use a regression model in terms of these parameters for the CEP tests. Possibly, the LRA tests are superior here due to their simplicity.

The score regression approach by Held et al. (2010) to test for ideal forecasters relies on independent forecast-observation tuples, and this restriction remains, when generalizing their approach to a test for cross-ideal forecasters. We present this test in Section 3 of the supplementary material. Finally, we remark that it is possible to derive a test for marginal cross-calibration by testing for mean zero in (6) for each $y \in \mathbb{R}$. It has turned out in simulations, that the resulting asymptotic test has several problems for applications. For completeness, we report these findings in Section 4 of the supplementary material.

7. Data example

The Bank of England (BoE) predicts the inflation rate of every quarter by using a probabilistic forecast with a potentially asymmetric two-piece normal distribution with parameters $\mu \in \mathbb{R}$, $\sigma_1, \sigma_2 > 0$ and density

$$f(y) = \begin{cases} \left(\frac{\pi}{2}\right)^{-1/2} (\sigma_1 + \sigma_2)^{-1} \exp\left\{-\frac{(y-\mu)^2}{2\sigma_1^2}\right\} & \text{if } y \leq \mu, \\ \left(\frac{\pi}{2}\right)^{-1/2} (\sigma_1 + \sigma_2)^{-1} \exp\left\{-\frac{(y-\mu)^2}{2\sigma_2^2}\right\} & \text{if } y > \mu. \end{cases} \tag{13}$$

The forecasts have been issued by the BoE's Monetary Policy Committee since February 1996 for the first quarter of 1996 and are publicly available online. The first quarter is from March to May, then from June to August, from September to November and the fourth quarter from December to February. Furthermore, there are forecasts available which have been issued between February 1993 and May 1997. These were converted into density forecasts retrospectively. Until the first quarter of 2004, the forecasts have been issued to predict RPIX inflation rates. But since the first quarter of 2004, inflation has been predicted and assessed in terms of percentage changes over twelve months of the CPI. The

observed RPIX as well as the CPI inflation rates are available from the Office for National Statistics under codes CDKQ and D7G7, respectively. There is no simple transformation that converts an RPIX inflation rate into a CPI inflation rate and vice versa, so we have analyzed the two data sets separately; RPIX inflation rate predictions from the first quarter of 1993 to the first quarter of 2004 and CPI inflation rate predictions from the first quarter of 2004 to the first quarter of 2015. In both cases we have 45 forecast-observation tuples. For further detail on the data set, see Gneiting and Ranjan (2011, Section 4.1). The BoE inflation forecasts have also been previously analyzed for example by Wallis (2003); Clements (2004); Mitchell and Hall (2005); Galbraith and van Norden (2012).

For both data sets, we compare the BoE predictions with a Gaussian autoregression (AR) of order one with rolling estimation window of length 20 quarters, which leads to Gaussian density forecasts. The prediction horizon we consider is one quarter. It is important to note that the AR forecast has the trivial σ-algebra $\{\emptyset, \Omega\}$ as additional information set $\mathcal{A}_{\mathrm{AR},t}$ as it only uses past realizations of inflation rates for prediction. Therefore, cross-calibration with respect to AR corresponds to cross-calibration with respect to $J = \emptyset$, which is the same as probabilistic calibration. As discussed in Section 6.3, the CEP and LRA tests make use of Assumption 2.13, which is trivially satisfied for the AR forecast and we believe it to be satisfied for the BoE forecasts.

First, we consider the CEP tests. The results for the BoE density forecasts can be seen in Figure 3 and the ones for the AR forecasts in Figure 4. In both plots the grid is $z_m = \{1 + (148/149)m\}/150$ for $0 \leq m \leq 149$ and $20'000$ bootstrap replications are used to calculate the adjusted p-values under the null hypothesis. For the RPIX inflation rate forecasts in the top panel of Figure 3, the BoE forecast seems to be probabilistically calibrated. According to the CEP test, it fails to be ideal, that is, cross-calibrated with respect to itself. The null hypothesis with respect to BoE is rejected for some small exceedance probabilities between zero and 0.05. However, the rejection region is small allowing the tentative conclusion that the BoE forecast is not far from being auto-calibrated. For the CPI inflation rate predictions in the bottom panel of Figure 3, the situation is different. Probabilistic calibration of the BoE forecast is rejected for exceedance probabilities between 0.13 and 0.26. Note that this result makes no use of Assumption 2.13. Cross-calibration with respect to BoE itself is also rejected in some parts of the region between 0.13 and 0.26.

According to the CEP test, the AR forecast for the RPIX inflation rate is not probabilistically calibrated with a minimal adjusted p-value of 0.029 and therefore also not cross-calibrated or cross-ideal; see the top panel of Figure 4. The overall conclusions remain the same for the CPI inflation rate forecasts; see the bottom panel of Figure 4. The minimal p-value for probabilistic calibration of the AR forecaster is 0.002, while it is 0.003 for cross-calibration with respect to BoE.

Secondly, we consider the LRA tests. The parametric class $\mathcal{F}$ used for the tests is the class of two-piece normal distributions with parameters $\mu \in \mathbb{R}, \sigma_1 > 0, \sigma_2 > 0$ given at (13). We can perform all the tests as for the CEP. The corre-

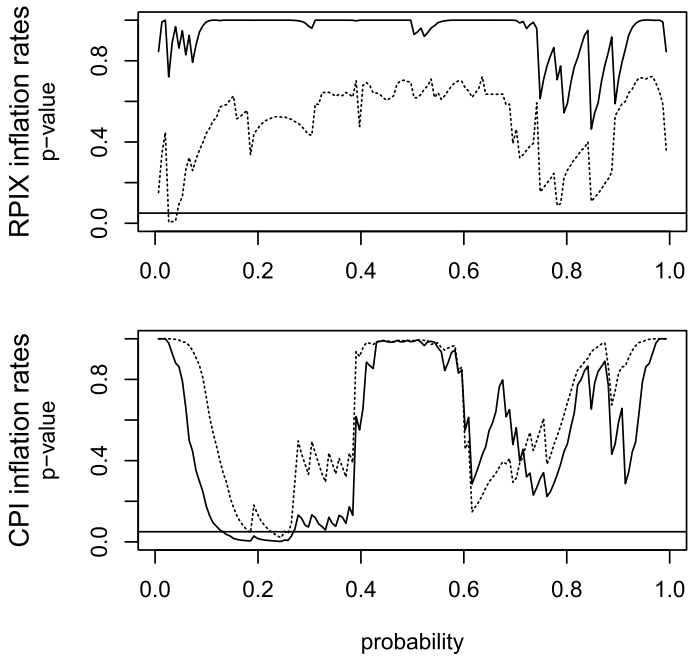

FIG 3. *The p-values of the CEP tests for the BoE forecast. The top panel corresponds to the prediction of RPIX inflation rates, whereas the bottom panel displays the results for CPI inflation rates. The solid horizontal lines give 0.05 level; the solid lines refer to probabilistic calibration and the dotted lines to cross-calibration with respect to BoE.*

TABLE 6
The p-values for the LRA tests for the BoE forecast.

	RPIX		CPI	
BoE wrt	$\emptyset$	BoE	$\emptyset$	BoE
F-test	[0.338]	0.010	[0.397]	0.149
AD-test	[0.496]	0.822	[0.010]	0.007
adjusted	[0.676]	0.021	[0.020]	0.015

TABLE 7
The p-values for the LRA tests for the AR forecast.

	RPIX		CPI	
AR wrt	$\emptyset$	BoE	$\emptyset$	BoE
F-test	0.0350	0.097	0.042	<0.001
AD-test	0.401	0.255	0.150	0.423
adjusted	0.070	0.195	0.084	<0.001

sponding p-values can be found in Tables 6 and 7. We also see if the estimated regression parameter failed to be zero or the standard normality assumption for the residuals was violated. Recall that the results in square brackets correspond to tests for probabilistic calibration in the case of a forecaster that uses a larger information set than $\mathcal{T}_t$. For the BoE forecaster, the overall results coincide with

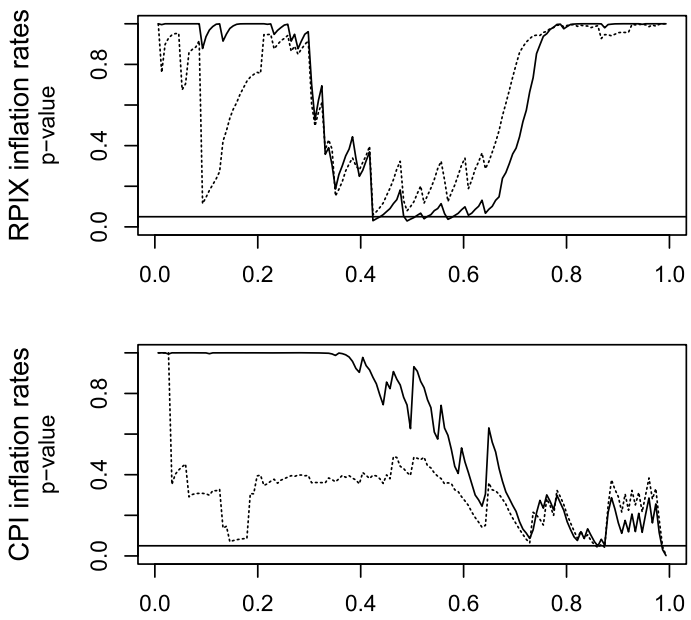

FIG 4. *The p-values of the CEP tests for the AR forecast. The top panel corresponds to the prediction of RPIX inflation rates, whereas the bottom panel displays the results for CPI inflation rates. The solid horizontal lines give the 0.05 level; the solid lines refer to probabilistic calibration and the dotted lines refer to cross-calibration with respect to BoE.*

the ones from the CEP. On the other hand, for the AR forecaster, the CEP tests display better power. Overall, for RPIX inflation rates, we would prefer the BoE forecast over the AR forecast, whereas for CPI inflation rates, the AR forecast seems to predict the lower tail of the distribution better than the BoE forecast. Therefore, one could investigate whether adapting the lower part of the two-piece normal distribution of the BoE forecast using the estimated parameters of the AR forecast improves calibration overall.

8. Discussion

We have extended the prediction space setting of Gneiting and Ranjan (2013) to accommodate serially dependent forecasts which are commonly encountered in practice. For prediction spaces with serial dependence, we have shown a refined version of the result of Diebold et al. (1998) on uniformity and independence of PIT values. It relies on Assumption 2.13, whose implications should be studied in greater detail. We have focussed on the case of one period ahead forecasts like in the original result. As mentioned in Remark 1, an analogous result continues to hold for q-step ahead forecasts. However, additional complications arise in testing for cross-calibration, which need further investigation in future research.

We have refined the notions of calibration to notions of cross-calibration

and we have provided powerful statistical tests for these properties requiring minimal assumptions on the sequences of forecasts and observations. The characterization of cross-calibration and cross-ideal forecasters in Proposition 2.11 sheds some light on the difference between ideal or auto-calibrated forecasters and probabilistically calibrated forecasters as discussed in Gneiting and Ranjan (2013). It is remarkable to note that with our approaches, testing for auto-calibrated forecasters is not more difficult than testing for probabilistic calibration.

However, in applications with strong temporal dependence of forecasts, it may be difficult to discern which information is purely derived from past observations of the quantity of interest, i.e. it belongs to $\mathcal{T}_t$, and which is additional information of a forecaster, i.e. it belongs to $\mathcal{A}_{i,t}$. This difficulty could possibly be overcome if the whole procedure for producing a forecast was disclosed by the forecaster.

The results of this paper allow to assess (cross-)calibration without the choice of a proper scoring rule, and hence, independently of sharpness considerations. Being aware of the calibration properties of a forecaster separately from its sharpness properties may provide guidance on how to improve forecasts and could potentially be useful in combining different forecasts.

In order to optimize forecasting performance, it is natural to combine forecasts. Gneiting and Ranjan (2013) have proposed combination formulas and aggregation methods to combine several forecasters; see also Ranjan and Gneiting (2010). It would be interesting to consider under which conditions calibrated forecasters can be combined to yield cross-calibrated forecasts. Also, the more refined notions of cross-calibration in this paper may help to identify which forecasters to include in combination formulas and which ones do not add additional information about the future outcome. Finally, combining forecasts is only a good idea if the predictions are based on different information sets. If there is a cross-calibrated forecaster with respect to all forecasters, any combination of forecasts would compromise on forecast quality.

Appendix A: Proofs of Section 2

Proof of Theorem 2.9. To show the first claim, observe that we have for all $y \in \mathbb{R}$,

$$\begin{aligned}\mathbb{E}_{\mathbb{Q}}\mathbb{1}\{F_{1,t}^{-1}(Z_{1,t}) \le y\} &= \mathbb{Q}[F_{1,t}(Y_{t+1}-)+V\{F_{1,t}(Y_{t+1}) - F_{1,t}(Y_{t+1}-)\} \le F_{1,t}(y)] \\ &= \mathbb{Q}\{F_{1,t}(Y_{t+1}) \le F_{1,t}(y)\} \\ &= \mathbb{Q}(Y_{t+1} \le y).\end{aligned}$$

The second equality holds, because

$$Z_{1,t} = F_{1,t}(Y_{t+1}-) + V\{F_{1,t}(Y_{t+1}) - F_{1,t}(Y_{t+1}-)\} \in [F_{1,t}(Y_{t+1}-), F_{1,t}(Y_{t+1})],$$

where the interval consists of the point $F_{1,t}(Y_{t+1}-) = F_{1,t}(Y_{t+1})$ if $F_{1,t}$ is continuous at the point Y_{t+1}, and $Z_{1,t} \in (F_{1,t}(Y_{t+1}-), F_{1,t}(Y_{t+1}))$ almost surely,

otherwise. Furthermore, $F_{1,t}(y) \leq F_{1,t}(Y_{t+1}-)$ or $F_{1,t}(y) \geq F_{1,t}(Y_{t+1})$. Let $J \subset \{i_1, \ldots, i_m\}$. The second claim follows because, for $y \in (0,1)$,

$$\begin{aligned}\mathbb{Q}(Z_{1,t} \leq y \mid F_{i,t}, i \in J, \mathcal{T}_t) &= \mathbb{E}_{\mathbb{Q}}\{\mathbb{Q}(Z_{1,t} \leq y | F_{i_1}, \ldots, F_{i_m}, \mathcal{T}_t) \mid F_{i,t}, i \in J, \mathcal{T}_t\} \\ &= \mathbb{E}_{\mathbb{Q}}(y \mid F_{i,t}, i \in J, \mathcal{T}_t) = y\end{aligned}$$

by the definition of cross-calibration. The last claim holds because

$$\mathbb{E}_{\mathbb{Q}} \mathbb{1}\{F_{2,t}^{-1}(Z_{1,t}) \leq y\} = \mathbb{E}_{\mathbb{Q}} \mathbb{Q}\{Z_{1,t} \leq F_{2,t}(y) \mid F_{2,t}, \mathcal{T}_t\} = \mathbb{E}_{\mathbb{Q}} F_{2,t}(y). \qquad \square$$

Proof of Proposition 2.11. The equivalence of parts one and two is immediate from the definition of cross-calibration. Suppose now that $1 \in \{i_1, \ldots, i_m\}$. For all $y \in \mathbb{R}$, we obtain

$$\begin{aligned}\mathbb{Q}(Y_{t+1} \leq y | F_{i_1,t}, \ldots, F_{i_m,t}, \mathcal{T}_t) &= \mathbb{Q}\{F_{1,t}^{-1}(Z_{1,t}) \leq y | F_{i_1,t}, \ldots, F_{i_m,t}, \mathcal{T}_t\} \\ &= \mathbb{Q}\{Z_{1,t} \leq F_{1,t}(y) | F_{i_1,t}, \ldots, F_{i_m,t}, \mathcal{T}_t\} = F_{1,t}(y),\end{aligned}$$

which shows the last claim. $\square$

Proof of Lemma 2.12. The forecast $F_{1,t}$ is $\sigma(F_{1,t})$ measurable and this is the smallest σ-algebra with this property. If $F_{1,t}$ is cross-ideal with respect to $\sigma(F_{2,t})$, then it is also measurable with respect to $\sigma(\sigma(F_{2,t}), \mathcal{T}_t)$ and hence $\sigma(F_{1,t}) \subset \sigma(\sigma(F_{2,t}), \mathcal{T}_t)$. Therefore, $\sigma(\sigma(F_{1,t}), \sigma(F_{2,t}), \mathcal{T}_t) = \sigma(\sigma(F_{2,t}), \mathcal{T}_t)$. $\square$

Proof of Theorem 2.14. We define the σ-algebra $\mathcal{B}_t := \sigma(\mathcal{A}_{i_1,t}, \ldots, \mathcal{A}_{i_m,t})$. For $u = (u_0, \ldots, u_l) \in (0,1)^{l+1}$, we obtain

$$\begin{aligned}&\mathbb{E}_{\mathbb{Q}} \left\{ \prod_{k=0}^{l} \mathbb{1}(Z_{1,t+k} \leq u_k) \mid \mathcal{B}_{t+l} \right\} \\ &\quad = \mathbb{E}_{\mathbb{Q}} \left[\prod_{k=0}^{l} \mathbb{E}_{\mathbb{Q}}\{\mathbb{1}(Z_{1,t+k} \leq u_k) \mid \mathcal{B}_{t+l}, \mathcal{T}_{t+k}\} \mid \mathcal{B}_{t+l} \right] \\ &\quad = \mathbb{E}_{\mathbb{Q}} \left[\prod_{k=0}^{l} \mathbb{E}_{\mathbb{Q}}\{\mathbb{1}(Z_{1,t+k} \leq u_k) \mid \mathcal{B}_{t+k}, \mathcal{T}_{t+k}\} \mid \mathcal{B}_{t+l} \right] = \prod_{k=0}^{l} u_k\end{aligned}$$

where we use condition (3) to obtain the second equality. $\square$

Appendix B: Proofs of Section 4

The proof of Theorem 4.1 parallels the proof of Gneiting and Ranjan (2013, Theorem 2.11). The following lemma gives a formula for the density function of Z_{p_1} conditional on $p_1 = x_1, \ldots, p_k = x_k$.

Lemma B.1. *The density function of Z_{p_1} conditional on $\mathbf{p} = \mathbf{x}$ is given by*

$$u(z | \mathbf{p} = \mathbf{x}) = \frac{q(\mathbf{x})}{x_1} \mathbb{1}(1 - x_1 \leq z \leq 1) + \frac{1 - q(\mathbf{x})}{1 - x_1} \mathbb{1}(0 \leq z < 1 - x_1),$$

where $\mathbf{p} = (p_1, \ldots, p_k)$, $\mathbf{x} = (x_1, \ldots, x_k) \in [0,1]^k$, $x_1 \in (0,1)$ and $q(\mathbf{x}) = \mathbb{Q}(Y = 1 | \mathbf{p} = \mathbf{x})$.

Proof of Lemma B.1. The PIT of p_1 is

$$Z_{p_1} = \begin{cases} (1-p_1) + p_1 V, & \text{if } Y = 1, \\ (1-p_1)V, & \text{if } Y = 0. \end{cases}$$

Let $0 \le z \le 1$, then

$$\begin{aligned}
\mathbb{Q}(Z_{p_1} \le z \,|\, \mathbf{p} = \mathbf{x}) &= \mathbb{Q}\{(1-p_1) + p_1 V \le z, Y = 1 | \, \mathbf{p} = \mathbf{x}\} \\
&\quad + \mathbb{Q}\{(1-p_1)V \le z, Y = 0 \,|\, \mathbf{p} = \mathbf{x}\} \\
&= \mathbb{Q}\{(1-p_1) + p_1 V \le z \,|\, Y = 1, \mathbf{p} = \mathbf{x}\}\mathbb{Q}(Y = 1 | \, \mathbf{p} = \mathbf{x}) \\
&\quad + \mathbb{Q}\{(1-p_1)V \le z \,|\, Y = 0, \mathbf{p} = \mathbf{x}\}\mathbb{Q}(Y = 0 | \, \mathbf{p} = \mathbf{x}) \\
&= \frac{z + x_1 - 1}{x_1} q(\mathbf{x}) \mathbb{1}(1 - x_1 \le z) + \{1 - q(\mathbf{x})\}\mathbb{1}(1 - x_1 \le z) \\
&\quad + \frac{z}{1 - x_1}\{1 - q(\mathbf{x})\}\mathbb{1}(1 - x_1 > z) \\
&= \frac{1 - q(\mathbf{x})}{1 - x_1} z \mathbb{1}(1 - x_1 > z) \\
&\quad + \left\{1 - \frac{q(\mathbf{x})}{x_1} + \frac{q(\mathbf{x})}{x_1} z\right\} \mathbb{1}(1 - x_1 \le z).
\end{aligned}$$

Differentiation yields the claim. □

Proof of Theorem 4.1. It is easy to see that part two is equivalent to part three. By Theorem 2.9, part three implies part one. The remaining task is to prove that part one implies part two. Let $\mathbb{H} = \mathbf{p}(\mathbb{Q})$ be the marginal law of the random vector $\mathbf{p}$ under $\mathbb{Q}$. Recall that $q(\mathbf{x}) = \mathbb{Q}(Y = 1 | \, \mathbf{p} = \mathbf{x})$. If $\mathbb{H}(\{0\} \times [0,1]^{k-1}) > 0$, then $q(0, x_2, \ldots, x_k) = 0$ for all $(x_2, \ldots, x_k) \in [0,1]^{k-1}$, because

$$\mathbb{H}(\{0\} \times [0,1]^{k-1}) = \mathbb{Q}\{\mathbf{p}^{-1}(\{0\} \times [0,1]^{k-1})\} = \mathbb{Q}\{p_1^{-1}(0)\} = \mathbb{Q}(p_1 = 0),$$

and furthermore,

$$\begin{aligned}
&\mathbb{Q}(Z_{F_1} = 1 \,|\, p_2 = x_2, \ldots, p_k = x_k) \\
&\quad \ge \mathbb{Q}(Z_{F_1} = 1, Y = 1, p_1 = 0 \,|\, p_2 = x_2, \ldots, p_k = x_k) \\
&\quad = \mathbb{Q}(Y = 1, p_1 = 0 \,|\, p_2 = x_2, \ldots, p_k = x_k) \\
&\quad = \mathbb{Q}(Y = 1 \,|\, p_1 = 0, p_2 = x_2, \ldots, p_k = x_k)\, \mathbb{Q}(p_1 = 0) \\
&\quad = q(0, x_2, \ldots, x_k)\, \mathbb{Q}(p_1 = 0).
\end{aligned}$$

We know that $\mathbb{Q}(Z_{p_1} = 1 | \, p_2 = x_2, \ldots, p_k = x_k) = 0$, because $\mathcal{L}(Z_{p_1} \,|\, p_2, \ldots, p_k)$ is standard uniform. This implies that $q(0, x_2, \ldots, x_k) = 0$. Similarly one can show that $\mathbb{H}(\{1\} \times [0,1]^{k-1}) > 0$ implies $q(1, x_2, \ldots, x_k) = 1$ for all $(x_2, \ldots, x_k) \in [0,1]^{k-1}$.

Using that $\mathcal{L}(Z_{p_1} | p_2, \ldots, p_k)$ is a standard uniform distribution and Lemma B.1, we have for a.a. $z \in (0,1)$, $\delta \in (0, 1-z)$

$$0 = u(z + \delta | \, p_2 = x_2, \ldots, p_k = x_k) - u(z | \, p_2 = x_2, \ldots p_k = x_k)$$

$$= \int_{[0,1]} \{u(z+\delta|\,\mathbf{p}=\mathbf{x}) - u(z|\,\mathbf{p}=\mathbf{x})\}\, d\mathbb{H}_1(x_1)$$
$$= \int_{[1-z-\delta,1-z)} \frac{q(\mathbf{x})-x_1}{x_1(1-x_1)}\, d\mathbb{H}_1(x_1),$$

where $\mathbb{H}_1 = p_1(\mathbb{Q})$ is the marginal law of p_1 under $\mathbb{Q}$. We define the signed measure μ for a given $(x_2,\dots,x_k) \in [0,1]^{k-1}$ as

$$\mu(A) = \int_A \frac{q(\mathbf{x})-x_1}{x_1(1-x_1)}\, d\mathbb{H}_1(x_1),$$

for all Borel sets $A \subset [a,b]$, where $0 < a < b < 1$. For $[c,d) \subset [a,b]$ we have shown before, that

$$\mu([c,d)) = \int_{[c,d)} \frac{q(\mathbf{x})-x_1}{x_1(1-x_1)}\, d\mathbb{H}_1(x_1) = 0.$$

Therefore, $\mu(B) = 0$ for all $B \in \mathcal{B}([a,b])$. In particular, $\{x_1 \in [a,b]|\, q(\mathbf{x}) > x_1\}$ and $\{x_1 \in [a,b]|\, q(\mathbf{x}) < x_1\}$ are $\mathbb{H}_1$ null sets and we have $q(\mathbf{x}) = x_1$ $\mathbb{H}_1$-a.s., hence,

$$p_1^{-1}\{q(\mathbf{x}) = x_1)\} = \{\omega : q\{p_1(\omega), x_2, \dots, x_k\} = p_1(\omega)\}$$
$$= \{\mathbb{Q}(Y = 1\,|\,p_1, p_2 = x_2, \dots p_k = x_k) = p_1\}$$

has $\mathbb{Q}$-measure 1 for all $(x_2,\dots,x_k) \in [0,1]^{k-1}$. Therefore, $\mathbb{Q}(Y=1|\,\mathbf{p}) = p_1$ $\mathbb{Q}$-a.s.. □

Appendix C: Calculations for Example 2.10

Let $\mu \sim \mathcal{N}(0,1)$ and let τ takes values 1 or -1 with equal probability independent of μ. Conditional on μ and τ, the observation is $Y \sim \mathcal{N}(\mu,1)$ and the forecasters have the following predictive distribution functions: $F_1(y) = \Phi(y-\mu)$, $F_2(y) = \Phi\left(y/\sqrt{2}\right)$, $F_3(y) = (1/2)\Phi(y-\mu) + (1/2)\Phi(y-\mu-\tau)$, $F_4(y) = \Phi(y+\mu)$ for $y \in \mathbb{R}$. As in Gneiting et al. (2007), we use the definitions $\Psi_+(x) = \frac{1}{2}\{\Phi(x) + \Phi(x-1)\}$, $\Psi_-(x) = \frac{1}{2}\{\Phi(x) + \Phi(x+1)\}$. Thus, $\Psi_-(x) = \Psi_+(x+1)$ and $\Psi_-^{-1}\{\Psi_+(x+1)\} = x$.

Proposition C.1. *The unfocused forecaster F_3 is cross-calibrated with respect to F_1, F_2, F_4.*

Proof. Let $y \in (0,1)$. We have

$$\mathbb{Q}(Z_{F_3} \le y|F_1,F_2,F_4) = \mathbb{Q}(Z_{F_3} \le y|\mu)$$
$$= \frac{1}{2}\mathbb{Q}\{\Psi_+(Y-\mu) \le y|\mu\} + \frac{1}{2}\mathbb{Q}\{\Psi_-(Y-\mu) \le y|\mu\}$$
$$= \frac{1}{2}\Phi\{\Psi_+^{-1}(y)\} + \frac{1}{2}\Phi\{\Psi_-^{-1}(y)\} = y.$$ □

Acknowledgements

The authors would like to thank Tilmann Gneiting for suggesting the data example and an anonymous reviewer for constructive comments. Calculations were performed on UBELIX (http://www.id.unibe.ch/hpc), the HPC cluster at the University of Bern.

Supplementary Material

Further Examples and the Score Regression Approach
(doi: 10.1214/17-EJS1244SUPPA; .pdf). We provide a short discussion of the cross-calibration test suggested by Feinberg and Stewart (2008) and give additional examples of diagnostic plots for cross-calibration. We generalize the test suggested by Held et al. (2010) to a test for cross-ideal forecasters. Finally, we discuss a natural approach for testing marginal cross-calibration, which, unfortunately, is useless in practice.

Computer Code
(doi: 10.1214/17-EJS1244SUPPB; .zip). The zip archive contains all R-code used in the paper and supplementary material.

References

Al-Najjar, N. I. and J. Weinstein (2008). Comparative testing of experts. *Econometrica 76*, 541–559. MR2406865

Anderson, T. W. and D. A. Darling (1954). A test of goodness of fit. *Journal of the American Statistical Association 49*, 765–769. MR0069459

Berkowitz, J. (2001). Testing density forecasts, with applications to risk management. *Journal of Business and Economic Statistics 19*, 465–474. MR1947080

Bröcker, J. (2009). Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society 135*(643), 1512–1519.

Campbell, S. D. (2005). A review of backtesting and backtesting procedures. *Finance and Economics Discussion Series, Federal Reserve 21*.

Clements, M. P. (2004). Evaluating the Bank of England density forecasts of inflation. *The Economic Journal 114*, 844–866.

Cox, D. D. and J. S. Lee (2008). Pointwise testing with functional data using the Westfall-Young randomization method. *Biometrika 95*, 621–634.

Dawid, A. P. (1984). Statistical theory: The prequential approach. *Journal of the Royal Statistical Society: Series A 147*, 278–290.

Dawid, A. P. (1985). Calibration-based empirical probability. *The Annals of Statistics 13*, 1251–1274.

Dawid, A. P. (1986). Probability forecasting. In S. Kotz, N. L. Johnson, and C. B. Read (Eds.), *Encyclopedia of Statistical Sciences*, Volume 7, pp. 210–218. Wiley, New York. MR0892738

Diebold, F. X., T. A. Gunther, and A. S. Tay (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review 39*, 863–883.

Feinberg, Y. and C. Stewart (2008). Testing multiple forecasters. *Econometrica 76*, 561–582.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika 80*, 27–38.

Galbraith, J. W. and S. van Norden (2012). Assessing gross domestic product and inflation probability forecasts derived from Band of England fan charts. *Journal of the Royal Statistical Society: Series A 175*, 713–727.

Gneiting, T., F. Balabdaoui, and A. E. Raftery (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B 69*, 243–268. MR2325275

Gneiting, T. and M. Katzfuss (2014). Probabilistic forecasting. *Annual Review of Stastistics and Its Application 1*, 125–151.

Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association 102*, 359–378.

Gneiting, T. and R. Ranjan (2011). Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business and Economic Statistics 29*, 411–422.

Gneiting, T. and R. Ranjan (2013). Combining predictive distributions. *Electronical Journal of Statistics 7*, 1747–1782.

Hamill, T. M. (2001). Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review 129*, 550–560.

Heinze, G., M. Ploner, D. Dunkler, and H. Southworth (2013). *logistf: Firth's bias reduced logistic regression*. R package version 1.21.

Heinze, G. and M. Schemper (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine 21*, 2409–2419.

Held, L., K. Rufibach, and F. Balabdaoui (2010). A score regression approach to assess calibration of continuous probabilistic predictions. *Biometrics 66*, 1295–1305.

Holzmann, H. and M. Eulert (2014). The role of the information set for forecasting – with applications to risk management. *The Annals of Statistics 8*, 595–621.

Knüppel, M. (2015). Evaluating the calibration of multi-step-ahead density forecasts using raw moments. *Journal of Business and Economic Statistics 33*, 270–281. MR3337062

Mason, S. J., J. Galpin, L. Goddard, N. Graham, and B. Rajartnam (2007). Conditional exceedance probabilities. *Monthly Weather Review 135*, 363–372.

Meinshausen, N., M. H. Maathuis, and P. Bühlmann (2011). Asymptotic optimality of the Westfall-Young permutation procedure for multiple testing under dependence. *The Annals of Statistics 39*(6), 3369–3391.

Mitchell, J. and S. G. Hall (2005). Evaluating, comparing and combining density forecasts using the KLIC with an application to the Bank of England and NIESR 'fan' charts of inflation. *Oxford Bulletin of Economics and Statistics 67S*, 995–1033.

Mitchell, J. and K. F. Wallis (2011). Evaluating density forecasts: Forecast combinations, model mixtures, calibration and sharpness. *Journal of Applied Econometrics 26*, 1023–1040. MR2843116

Montgomery, D. C., E. A. Peck, and C. G. Vining (2001). *Introduction to linear regression analysis* (3rd ed.). John Wiley & Sons, Inc.

Murphy, A. H. (1994). A coherent method of stratification within a general framework for forecast verification. *Monthly Weather Review 123*, 1582–1588.

Murphy, A. H. and R. L. Winkler (1987). A general framework for forecast verification. *Monthly Weather Review 115*, 1330–1338.

Murphy, A. H. and R. L. Winkler (1992). Diagnostic verification of probability forecasts. *International Journal of Forecasting 7*, 435–455.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Ranjan, R. and T. Gneiting (2010). Combining probability forecasts. *Journal of the Royal Statistical Society: Series B 72*, 71–91.

Rüschendorf, L. (2009). On the distributional transform, Sklar's theorem, and the empirical copula process. *Journal of Statistical Planning and Inference 139*, 3921–3927.

Shapiro, S. S. and M. B. Wilk (1965). An analysis of variance test for normality (complete samples). *Biometrika 52*, 591–611.

Strähl, C. and Ziegel, J. (2017a). Further Examples and the Score Regression Approach. DOI: 10.1214/17-EJS1244SUPPA.

Strähl, C. and Ziegel, J. (2017b). Computer Code. DOI: 10.1214/17-EJS1244SUPPB.

Tsyplakov, A. (2011). Evaluating density forecasts: A comment. *MPRA paper 31233*. (Available from http://mpra.ub.uni-muenchen.de/31233).

Tsyplakov, A. (2013). Evaluation of probabilistic forecasts: proper scoring rules and moments. *Preprint*, http://dx.doi.org/10.2139/ssrn.2236605.

Wallis, K. F. (2003). Chi-squared tests of interval and density forecasts, and the bank of england's fan charts. *International Journal of Forecasting 19*, 165–175.

Westfall, P. and S. Young (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*. Wiley, New York.

Yap, B. W. and C. H. Sim (2011). Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation 81*, 2141–2155.